2016

Genomic prediction using haplotypes in New Zealand dairy cattle

Melanie Kate Hayr
Iowa State University

Genomic prediction using haplotypes in New Zealand dairy cattle

by

Melanie Kate Hayr

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Animal Breeding and Genetics (Quantitative Genetics)

Program of Study Committee:
Dorian Garrick, Major Professor
Jack Dekkers
Rohan Fernando
Alicia Carriquiry
Jarad Niemi

Iowa State University

Ames, Iowa

2016

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would first like to thank my major professor Dr Dorian Garrick for his guidance, patience and support throughout my Masters and PhD degrees – thank you for all the opportunities you have provided for me and your valuable insights into my work. Thank you to my committee members: Dr Jack Dekkers, Dr Rohan Fernando, Dr Alicia Carriquiry and Dr Jarad Niemi; for always being willing to offer their advice and encouragement, both inside and outside of the classroom. Thank you to Livestock Improvement Corporation, in particular Dr Richard Spelman and Dr Bevin Harris, for supporting me throughout my PhD and allowing me the opportunity to study at such a prestigious university.

I would like to thank past and present members of Iowa State University's Animal Breeding and Genetics group for being such a supportive and engaging group that allowed me to grow both personally and professionally. I would particularly like to thank my officemates: Jenn, Tracy, Brittany, Jenelle, Lydia, Emily, Laura and Jessie – you are such a wonderful group of strong women that will go so far in life! Thank you to all the Animal Science graduate students, faculty and staff for creating such a positive and illustrious work atmosphere.

Thanks to all my wonderful friends that have supported me and helped me grow to be the person I am today. Thank you to Cat, Claire and Ben for always being there through everything and waiting patiently for me to return to New Zealand. Thank you to Angelica for sharing all the experiences on the journey through graduate school and for travelling all the way to New Zealand for my wedding. Thanks to Justin, Junmarie, Johed, Emily, Adam and

Ben for all the good times and comic relief. Thank you to Kurt for always being there when Andrew or I needed you – you're awesome, dude.

I would like to thank my family – in both New Zealand and America. Thanks to my mum and dad for always supporting me in my endeavors. Thanks to Tania and Bryce for always challenging and supporting me. Thank you to Bret, Carolyn, Oma, Papa and Andrew's extended family for making me feel so welcome and included in your family. Thanks to Arty for always making me laugh. Finally I would like to thank my husband Andrew, for putting up with me and always being willing to talk science. I couldn't imagine this journey without you; I look forward to moving to New Zealand with you next year!

ABSTRACT

The genetic merit of livestock is now routinely evaluated using SNP genotype information on selection candidates to improve genetic gain. Improvement in genomic prediction accuracy will have a direct impact on genetic gains. The New Zealand dairy cattle population contains two major breeds: Holstein Friesian and Jersey; and KiwiCross, their admixed descendants are popular with many farmers. Genomic prediction models fitting haplotype alleles rather than SNPs have been shown to increase genomic prediction accuracy in simulated and purebred populations, but has not been assessed in admixed populations. This dissertation investigated whether prediction accuracy, and thus rate of genetic gain, can be improved in the admixed New Zealand dairy cattle population by fitting covariates for haplotype alleles rather than covariates for SNPs in genomic prediction models. Haplotype alleles were constructed from the phased genotypes at ~40,000 SNPs for ~58,000 New Zealand dairy cattle.

A measurement of the reliability of genomic prediction accuracy estimates is important for evaluating whether the performance of different genomic prediction models is significantly different. Chapter IV explored a method for calculating the posterior distribution of prediction accuracy from Bayesian genomic prediction models from calculating prediction accuracy in each iteration of the post-burn-in Markov chain Monte Carlo chain. Using 200 replicates of a simulated data set of 5,000 training and 1,000 validation individuals genotyped at 20,000 SNPs, our method appropriately captured the confidence in accuracy between true and estimated genetic merit but not between phenotype and estimated genetic merit. In practice the true genetic merit is not observed so the accuracy

between true and estimated genetic merit cannot be calculated. Further research is needed to assess the reliability of prediction accuracy estimates when true genetic merit is unknown.

The use of genomic prediction models that fit covariates for haplotype alleles, which were constructed based on a fixed length (e.g. 250 kb) or based on a population parameter (e.g. recombination), has potential for increasing genetic gain in admixed populations compared to fitting covariates for SNPs. The best model explored for this data set was based on recombination (up to 7.7% improvement in prediction accuracy over the SNP model; $p < 0.001$); however, the best method for assigning SNPs to haplotype blocks may differ in admixed populations compared to purebred populations because patterns of linkage disequilibrium may be different between breeds within the population. The best method will also depend on the number of individuals genotyped and the relationships between them. Consistent with results in other populations, fixed-length haplotypes appear to perform well in the New Zealand dairy cattle population (up to 5.5% improvement over the SNP model; $p = 0.002$) as long as haploblock length is appropriate. Haplotype blocks generated using recombination events within the population may provide higher prediction accuracy if these recombination events can be accurately identified (i.e. many closely related animals). Removal of rare haplotype alleles from the data set reduces the computational demands of genomic prediction fitting haplotype alleles with no loss in prediction accuracy. Further reduction in computational demands and improvement in prediction accuracy could be obtained by fitting combined SNP and haplotype models. The increase in the number of individuals genotyped and sequenced will likely improve the benefits of fitting haplotype alleles in genomic prediction models.

CHAPTER I

GENERAL INTRODUCTION

1.1 Introduction

Animal breeding programs are designed to make progress towards a specified breeding objective. An individual's phenotype is impacted by two factors: the genetic merit of that individual and environmental (non-genetic) factors (e.g. age, diet, management). The breeding value (BV) of an individual is the additive genetic portion of an individual's phenotype that can be passed to its offspring. Sustained population-wide progress towards the breeding objective can be made when individuals that are selected as parents have superior BV compared to the population as a whole. It is not possible to directly observe the BV of an individual so it must be estimated through the fitting of a statistical model to obtain an estimated breeding value (EBV). The rate of progress in the breeding objective can be predicted as:

$$\Delta g = \frac{i r_{EBV,BV} \sigma_A}{L} \qquad [1.1]$$

where Δg is the expected genetic gain, i is the selection intensity in standard deviations, r is the estimated accuracy of selection, σ_A is the estimated additive genetic standard deviation of the population for the selection criteria and L is the generation interval (Lush, 1937). Equation 1.1 has been expanded to allow for different selection pathways in males and females (Falconer and Mackay, 1996), and also to the four-path system that is common in dairy cattle, encompassing sires of sires, dams of sires, sires of dams and dams of dams (Rendel and Robertson, 1950).

Most features in equation 1.1 are under biological constraints: the genetic standard deviation is not easily manipulated and the generation interval, defined as the average age of individuals in the population when their offspring are born, cannot be reduced below the time it takes to reach sexual maturity plus gestation length. Selection intensity is determined by the proportion of animals selected to be parents and is limited by the number of individuals that need to be bred as replacement animals, as well as the biological limitation of how many offspring an individual can have in a given timeframe and limits on the rate of inbreeding. The final feature in this equation is the prediction accuracy, which is influenced by the individuals that have recorded information, the heritability of the trait, the statistical model used to obtain EBV and the genetic make-up of the population (e.g. single-breed or mixed-breed, family structure) (Hayes et al., 2010). The implementation of statistical models with improved prediction accuracy will improve the rate of genetic gain in the population.

Traditionally, pedigree-based Best Linear Unbiased Prediction (BLUP) models have been used to obtain EBV (Henderson, 1984). Pedigree BLUP accounts for the expected genetic relationship between individuals based on pedigree information (Wright, 1922). The development of single nucleotide polymorphism (SNP) panels for livestock improved estimates of the genetic relationship between individuals due to the ability to resolve pedigree-errors, and also capture similarity between closely-related animals due to Mendelian sampling (Van Raden, 2008). Models that account for the relationship between individuals using genetic markers are termed genomic-BLUP (GBLUP) models (Van Raden, 2008). A variety of marker effects models, whereby effects are estimated for each SNP, have also been explored (Meuwissen et al., 2001; Kizilkaya et al., 2010; Habier et al., 2011). Models using SNPs to obtain EBV could only be used to evaluate the genetic merit of genotyped individuals; therefore two-step approaches

were needed to first estimate the genetic merit of genotyped individuals, which were then blended with traditional PBLUP evaluations (Harris and Johnson, 2010). Recently, methods have been described to combine pedigree and genotype information to allow the estimation of breeding values for all individuals from a single model (Harris et al., 2012; Fernando et al., 2014).

Current research in genomic prediction focuses on further improving prediction accuracy using multiple approaches, including the development of Bayesian models with different prior assumptions, increasing marker density up to whole-genome sequence, or identifying and modeling context-specific QTL, e.g. genotype by environment interactions. The focus of this literature review will be to discuss the current state of genomic prediction, including advancements and limitations, and discuss the potential of fitting haplotype alleles rather than SNPs in genomic prediction models.

1.2 Biology of Inheritance

Dairy cattle have 29 pairs of autosomes and a single pair of sex chromosomes (X and Y), totaling 60 chromosomes, with one copy of each autosome and one sex chromosome inherited from each parent. Meiosis is the process by which a single diploid cell ($n_{chr} = 60$) divides into four haploid gametes ($n_{chr} = 30$) (Figure 1.1; Marston and Amon, 2004). The first step of meiosis is each pair of homologous chromosomes align with each other and replication occurs to result in two copies of each of the chromosomes ($n_{chr} = 120$). Recombination events occur between one maternally-inherited and one paternally-inherited chromosome through the formation of a Holliday junction at a chiasma. After recombination occurs, the chromatids independently separate into four haploid gametes ($n_{chr} = 30$).

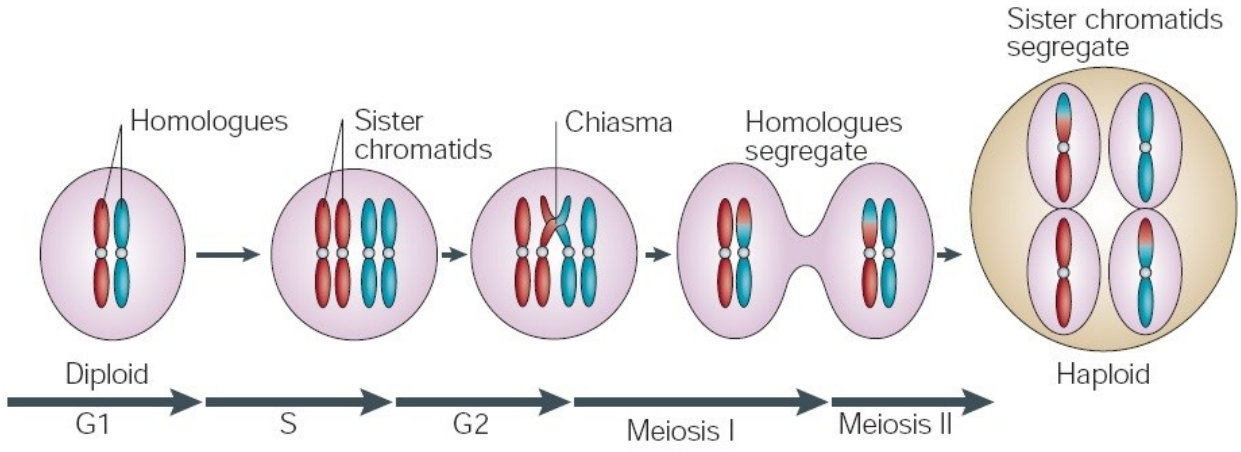


Figure 1.1: The Meiotic Cell Cycle. A diploid cell contains two copies of each autosomal chromosome: one inherited from the mother (red) and one inherited from the father (blue). Recombination events can occur at a chiasma between homologous chromosomes. Figure modified from Marston and Amon (2004).

The independent separation of each chromosome results in a 50% chance that two alleles at loci on different chromosomes will be present in the same gamete. The chance that two alleles at loci on the same chromosome will be present in the same gamete depends on how often there is a recombination event between the two loci. A morgan (M) is a measurement of the expected number of recombinations during a cycle of meiosis; the dairy cattle genome is approximately 30 M, and autosomal chromosome length ranges from 57 to 166 centimorgans (cM; Arias et al., 2009). The distance between two loci in centimorgans is related to the physical map distance between the loci in megabases (Mb), and on average there is 1.25 cM per Mb in New Zealand dairy cattle (Arias et al., 2009); therefore alleles at loci that are close together on a chromosome are more likely to be inherited together than alleles at loci that are further apart on the chromosome. Linkage disequilibrium (LD) refers to the non-random association of alleles at different loci. Factors such as genetic drift, migration, selection, mutation and population bottlenecks generate LD; while recombination breaks down LD. Genomic prediction using a

marker effects model takes advantage of LD between the (typically unobserved) causal mutations (QTL) and genotyped SNP markers to capture the effects of the QTL (Dekkers, 2007).

1.3 Genomic Prediction

Genomic Prediction Models

Breeding Value Models

The traditional method for estimating breeding values directly estimates the breeding values, fitted as random effects, based on the expected correlation between the BV of relatives due to pedigree relationships:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \qquad [1.2]$$

where **y** is a vector of phenotypes, **X** and **Z** are incidence matrixes, **u** is a vector of breeding values and **e** is a vector of residuals. The EBV can be obtained for pedigree-based BLUP by solving the Mixed Model Equations (Henderson, 1984):

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{A^{-1}\lambda} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix} \qquad [1.3]$$

where $\lambda = \sigma_e^2 / \sigma_u^2$ and $\boldsymbol{A}^{-1}$ is the inverse of the pedigree-based relationship matrix. Pedigree-based BLUP uses the expected relationships (i.e. covariances) between individuals, using the probability that for a given locus, the allele present is identical by descent, i.e. were inherited from the same common ancestor (Wright, 1922) to estimate breeding values. Two full-siblings are expected to share half their alleles because each of them inherits half the genes from each parent. The half of each parents' genome that each sibling inherits will vary due to recombination events, the independent assortment of chromosomes during gamete formation and the random selection of gametes in the formation of the embryo. This phenomenon is called

Mendelian sampling and the reason why some pairs of siblings will have more than half of their alleles in common and others will have less than half. Comparison of the SNP genotypes that two individuals share provides information on the covariance between individuals at the genomic level (Nejati-Javaremi et al, 2007). A common method to calculate the relationship between individuals based on genotype information, referred to as the genomic relationship matrix, was described by Van Raden (2008):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{\sum_i 2p_i q_i} \qquad [1.4]$$

where **Z** is the column-centered matrix **M** of animals and SNP genotypes, p_i is the frequency of one allele at SNP i and q_i is the frequency of the other allele. An example of an **M** matrix, showing the count of a particular allele for each individual at each SNP, is in Table 1.1. Mixed model equations that replace $\boldsymbol{A}^{-1}$ in Equation 1.3 with the inverse of the **G** matrix ($\boldsymbol{G}^{-1}$) are referred to as genomic BLUP (GBLUP) (VanRaden, 2008) and have been shown to improve prediction accuracy over pedigree-BLUP models (Hayes et al., 2009a; Wolc et al., 2011b; Daetwyler et al., 2012).

Table 1.1: Example of a Genotype Matrix for 5 Animals and 9 SNPs

ID	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
1001	0	1	2	1	0	1	1	2	2
1002	2	1	0	2	2	0	2	1	1
1003	1	1	2	1	0	2	0	0	2
1004	1	1	0	1	2	0	1	1	2
1005	1	2	1	1	1	1	0	1	1

Marker Effects Models

Marker effects models directly estimate effects for each SNP in the **M** matrix rather than breeding values for each individual.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{M\alpha} + \boldsymbol{e} \qquad [1.5]$$

where **y, X, M, β and e** are as described for equations 1.2 and 1.3, **α** is a vector of the random effect of each SNP, and the EBV ($\widehat{\boldsymbol{u}}$) (or Direct Genomic Values (DGV)) are $\mathbf{M}\widehat{\boldsymbol{\alpha}}$. Bayesian marker effects models were first described by Meuwissen et al. (2001) and require prior assumptions to be made about the distribution of **α, β,** and **e**. Meuwissen et al. (2001) described two marker effects models, BayesA and BayesB. BayesA assumes that SNP effects have identical and independent t-distributions with scale parameter S_α^2 and ν_α degrees of freedom. This prior is equivalent to assuming SNP effects have identical and independent Normal distributions with a mean of zero and locus-specific variance (i.e. $\alpha \sim N(0, \sigma_{\alpha_j}^2)$) (Fernando and Garrick, 2013). BayesB assumes SNP effects have a mixture distribution such that:

$$\alpha_i \sim \begin{cases} t(\nu_\alpha, S_\alpha^2) & \text{with probability } 1-\pi \\ 0 & \text{with probability } \pi \end{cases} \qquad [1.6]$$

where π is the proportion of SNPs that are assumed to have an effect of zero.

There have since been many modifications to their methods, focused primarily on the prior assumptions for the SNP effects (**α**) (Gianola et al., 2009; Habier et al., 2011; Gianola, 2013). BayesC (Kizilkaya et al., 2010) is similar to BayesB but instead of assuming marker effects have identical and independent t-distributions, the assumption is that they have identical and independent Normal distributions with a variance that is common across all markers:

$$\alpha_i \sim \begin{cases} N(0, \sigma_\alpha^2) & \text{with probability } 1-\pi \\ 0 & \text{with probability } \pi \end{cases} \qquad [1.7]$$

When π = 0 and the variance parameters are not sampled, this model is equivalent to the breeding-value GBLUP model (Stranden and Garrick, 2009; Shen et al., 2013). An extension to this model by Habier et al. (2011) describes the model BayesCπ, in which π is treated as unknown with a Uniform(0,1) prior.

Methods BayesA, B and C all assume that SNP effects are independent; however the effects of SNPs surrounding a QTL are unlikely to be independent. BayesN was described by Zeng (2015) to allow for the dependence of SNPs that are closely linked. In BayesN, SNPs are clustered into windows of neighboring SNPs based on their position along the genome and assumes that some windows are not associated with the trait (and SNPs have an effect of zero). Π is the probability that a window has an effect of zero and π_j is the window-specific probability that a SNP in window j has an effect of zero. Marker effects are assumed to have a window-specific variance:

$$\alpha_{ij} \sim \begin{cases} N\left(0, \sigma^2_{\alpha_j}\right) & \text{with probability } (1-\Pi)\left(1-\pi_j\right) \\ 0 & \text{else} \end{cases} \quad [1.8]$$

No genomic prediction model outperforms the other models in all situations (Daetwyler et al., 2013). The most appropriate model for genomic prediction has been shown to depend on the trait heritability and genetic architecture (Daetwyler et al., 2010; Hayes et al., 2010), relationships between genotyped individuals (Habier et al., 2007) and the number of individuals with both genotypes and phenotypes (de los Campos et al., 2013), among other things (Daetwyler et al., 2013). BayesA and BayesB can provide higher prediction accuracy than other methods when the trait is controlled by large QTL (Meuwissen et al., 2001). BayesC shrinks large effects more than BayesA or BayesB (Kizilkaya et al., 2010; Gondro et al., 2013) and is less sensitive to priors for genetic and phenotypic variances (Habier et al., 2011). BayesN has been shown to capture effects of low-frequency QTL better than BayesB (Zeng, 2015). It is therefore common to explore the performance of a range of models prior to implementation in a genomic selection program.

Current State of Genomic Prediction Models

Genomic selection has been successfully implemented in plant and animal species around the world (Hill, 2014) and has been particularly successful in dairy cattle because of the reduction of generation intervals as a result of higher-accuracy EBV being available at a young age (Schaeffer, 2006; Hayes et al., 2009b). Genomic prediction models take advantage of markers that are in LD with QTL to capture the effects of those QTL (Dekkers, 2007), but also capture relationships between individuals (Habier et al., 2007). It has been hypothesized that increasing the number of genotyped SNPs across the genome would increase prediction accuracy (Meuwissen and Goddard, 2010); however increasing the number of SNPs from ~50,000 to ~700,000 has resulted in only slight increases in accuracy (Su et al., 2012; Erbe et al., 2014). It was then hypothesized that whole-genome sequence was required to improve prediction accuracy substantially (Meuwissen et al., 2013); however, again, this has not led to an improvement in prediction accuracy (van Binsbergen et al., 2015; Heidaritabar et al., 2016). It is likely that the lack of improvement in accuracy is because there are many more SNPs than individuals with genotypes, therefore the effects of most SNPs are shrunken too much (Gianola, 2013). Imputation from SNP panels to sequence, discussed in detail below, was required to obtain enough individuals with sequence information for genomic prediction (van Binsbergen et al., 2015; Heidaritabar et al., 2016) and errors in that process may also contribute to the lack of improvement in accuracy.

A variety of other approaches are currently being explored to improve genomic prediction accuracy. Sequence data is being explored to identify QTL in an attempt to decrease the number of markers that are fitted in genomic prediction models (MacLeod et al., 2016). Genomic prediction models that use SNP genotype information to impute putative QTL have

also been explored (Zeng, 2015). Models that attempt to identify or capture context-dependent QTL effects may also improve prediction accuracy, e.g. QTL that are dependent on breed of origin (Saatchi et al., 2014), parent of origin (Tuiskula-Haavisto et al., 2004) or environment to which the animal is exposed (Mulder, 2016).

1.4 Haplotype Models

Recombination Events

Recombination events do not occur randomly along the genome: there are recombination hotspots where recombination occurs at a higher frequency than average, and recombination coldspots where recombination occurs at a lower than average frequency (Sandor et al., 2012; Weng et al., 2014). Recombination acts to generate new combinations of alleles, increasing the genetic diversity of the population, which may lead to an evolutionary advantage. However, recombination events can cause instability in the genome or break up favorable combinations of alleles; therefore, recombination hotspots and coldspots may have evolved to limit these negative consequences of recombination (Kauppi et al., 2004).

Patterns of recombination across the genome have been studied in many species including yeast (Baudat and Nicolas, 1997), mice (Paigen and Petkov, 2010), humans (Jeffreys et al., 2005), cattle (Sandor et al., 2012; Weng et al., 2014), and *Arobidopsis thaliana* (Mezard, 2006). Recombination hotspots have been shown to occur near genes but recombination frequencies within genes are often low (Myers, 2005). The location of recombination hotspots has been shown to change over time, and depends on the sex of the individual (Paigen and Petkov, 2010; Kauppi et al., 2004; Jeffreys et al., 2005). Epigenetic features such as DNA methylation and histone folding play a role in the location of recombination events, e.g. DNA

that is tightly folded is less likely to be able to form Holliday Junctions (Robertson and Wolffe, 2000; Borde et al., 2009).

The LD between markers that surround a recombination hotspot is likely to be lower than between markers that occur in a recombination coldspot. Haploblocks are genomic regions comprising neighboring genetic markers whose phased alleles are likely to be inherited together (i.e. with few recombination events between them). A haplotype allele is the combination of phased SNP alleles that are present in a haploblock (Figure 1.2). Estimating effects for haplotype alleles rather than SNPs in genomic prediction models may increase prediction accuracy for multiple reasons, described later in this section.

Genotype Phasing

The current SNP panel approach to genotyping individuals does not provide information as to which allele was inherited from the sire vs. the dam, which is information that is required to identify haplotype alleles. Evaluation of the genotypes across individuals can provide information about which alleles at neighboring markers were inherited from the same parent (Browning and Browning, 2011). Some individuals may have missing genotypes for some markers and information about the phased alleles at neighboring loci can provide reliable information as to the genotype of the individual at the missing marker, which is known as imputation.

The accuracy of phasing and imputation methods depends on the size of the data set, the relationship between individuals and the method used for phasing. Larger data sets of individuals usually have higher accuracy because a greater proportion of haplotypes in the population are captured. Phasing accuracy is higher in data sets where individuals have multiple relatives in the

data set compared to when individuals are more distantly related (Browning and Browning, 2011; Weng et al., 2014; Ferdosi et al., 2016).

Two types of information can be taken into account when phasing: population-wide haplotypes and pedigree information. Programs such as BEAGLE have traditionally only taken into account population-wide haplotypes (Browning and Browning, 2007) but have more recently enabled the incorporation of pedigree information (Browning and Browning, 2009). LINKPHASE3 (Druet and Georges, 2015) initially uses pedigree information to phase genotypes, however it may be difficult to resolve phase in some regions based on pedigree information so DAGPHASE (Druet and Georges, 2010) utilizes population-wide haplotypes from BEAGLE to improve phasing accuracy in these regions. Some phasing programs output predicted haploblocks based on recombination or LD patterns observed in the data set (Crawford et al., 2004; O'Connell et al., 2014; Druet and Georges, 2015); however the confidence in these haploblocks depends on the ability to accurately phase the data and therefore relies on size of the data set and relationships between individuals.

Genomic Prediction Fitting Haplotypes

After phasing genotypes into haplotypes there are three steps to fitting haplotype alleles in genomic prediction models: 1) defining haploblocks; 2) generating the haplotype matrix for those haploblocks and 3) running the genomic prediction model. Genomic prediction fitting haplotypes has been explored in simulated (Villumsen and Janss, 2009; Villumsen et al., 2009) and real (Hayes et al., 2007) data; however the focus has been on purebred populations, ignoring admixed-breed populations.

Genotype Matrix

ID	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
1001	0	1	2	1	0	1	1	2	2
1002	2	1	0	2	2	0	2	1	1
1003	1	1	2	1	0	2	0	0	2
1004	1	1	0	1	2	0	1	1	2
1005	1	2	1	1	1	1	0	1	1

Phased Haplotypes

ID	Strand	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
1001	Sire	0	1	1	0	0	1	1	1	1
1001	Dam	0	0	1	1	0	0	1	1	1
1002	Sire	1	1	0	1	1	0	0	0	1
1002	Dam	1	0	0	1	1	0	1	1	0
1003	Sire	0	1	1	1	0	1	0	0	1
1003	Dam	1	0	1	0	0	1	1	1	1
1004	Sire	0	1	0	0	1	0	1	1	1
1004	Dam	1	0	0	1	1	0	0	0	1
1005	Sire	0	1	1	0	0	1	0	0	1
1005	Dam	1	1	0	1	1	0	0	1	0

Let SNP3-SNP6 belong to Haploblock A:

ID	SNP3	SNP4	SNP5	SNP6
1001	1	0	0	1
1001	1	1	0	0
1002	0	1	1	0
1002	0	1	1	0
1003	1	1	0	1
1003	1	0	0	1
1004	0	0	1	0
1004	0	1	1	0
1005	1	0	0	1
1005	0	1	1	0

ID	Haplotype Alleles
1001	1001
1001	1100
1002	0110
1002	0110
1003	1101
1003	1001
1004	0010
1004	0110
1005	1001
1005	0110

Unique Alleles	
Allele1	1001
Allele2	1100
Allele3	0110
Allele4	1101
Allele5	0010

Haplotype Matrix for Haploblock A

ID	Allele1	Allele2	Allele3	Allele4	Allele5
1001	1	1	0	0	0
1002	0	0	2	0	0
1003	1	0	0	1	0
1004	0	0	1	0	1
1005	1	0	1	0	0

Figure 1.2: Generating the Haplotype Matrix. Genotypes are phased to identify alleles that were inherited from the same parent. SNP alleles in a haploblock are combined to generate the haplotype alleles. Unique haplotype alleles are identified. The haplotype matrix is generated that contains a column for each unique haplotype allele and the 0/1/2 count of the number of copies of that haplotype allele that are present in that individual.

Multiple approaches have been described to obtain haploblocks, each with its own strengths and weaknesses. Haploblocks of a fixed length throughout the genome, using either SNPs (Hayes et al., 2007; Calus et al., 2009; Villumsen et al., 2009) or Megabases (Sun et al., 2016), are computationally easy to implement, but they do not take population-specific information on LD or recombination events into account, and therefore recombination events may frequently occur within a given haploblock. Alternative approaches, termed variable-length methods, attempt to more closely capture the co-segregation of alleles within the population through identification of recombination hotspots, measurement of LD, or reducing the number of haplotype alleles that are present in a population (Rinaldo et al., 2005). Although these approaches may capture co-segregation of alleles more accurately, they are much more time-consuming to generate and haploblocks are specific to the population, which is not ideal for genomic prediction across populations. A number of programs exist to predict haploblock structure in the population, e.g. Haploview (Barrett et al., 2005) and CFHLC+ (Mourad et al., 2011); however many of these programs are optimized for fewer individuals than is needed for genomic prediction or for sequence information on a small genomic region and are not feasible to use on a genome-wide, or chromosome-wide level.

The process for generating the haplotype matrix for a single haploblock is outlined in Figure 1.2. The matrix is constructed for each haploblock and horizontally concatenated to provide a matrix with the number of rows equal to the number of individuals and the number of columns equal to the number of haplotype alleles across the whole genome.

Performance of Haplotype Models

Genomic prediction studies that fitted haplotype alleles have observed an improvement in prediction accuracy over fitting SNPs but that prediction accuracy decreased if haploblock length was too long or too short (Calus et al., 2008; Calus et al., 2009; Villumsen and Janss, 2009; Villumsen et al., 2009; Boichard et al., 2012). Most of these studies evaluated haplotypes that were generated using only one method (i.e. fixed length or a particular variable length method) and compared performance to that of the SNP model. These studies also primarily focused on simulated data sets, data sets from pure-bred populations, or data sets with closely related breeds, and therefore may not translate to admixed populations with diverse breeds.

Possible Reasons for Improved Accuracy

The improved accuracy that is typically observed when fitting covariates for haplotype alleles rather than SNPs could be due to a combination of: 1) improved ability to detect ancestral relationships (Ferdosi et al. 2016); 2) higher LD between QTL and haplotypes than QTL and SNPs (Zondervan and Cardon, 2004); and 3) the ability to capture short-range epistatic effects (e.g. Littlejohn et al., 2014).

The **A** matrix is calculated based on the expected relationship between individuals based on pedigree information and captures the expected probability that alleles between two individuals are identical-by-descent (IBD; Wright, 1922). The **G** matrix is calculated based on whether the same SNP allele is present in the two individuals, based on SNPs selected for a SNP panel (Zondervan and Cardon, 2004). The **G** matrix may generate non-zero relationships between unrelated individuals (i.e. individuals of different breeds; Harris and Johnson, 2010) because it is reliant on genotype states rather than IBD. The relationships from the **G** matrix typically have a different mean and variance for both diagonal and off-diagonal elements than

the **A** matrix generated for the same set of animals (Forni et al., 2011). Hickey et al. (2013) and Ferdosi et al. (2016) showed that combinations of phased SNP alleles (i.e. haplotype alleles) capture true IBD relationships better than SNP genotypes. Therefore the use of haplotype alleles can lead to improved prediction accuracy over either pedigree-based or SNP-based genomic prediction because they capture IBD relationships more accurately than SNP genotypes but also capture Mendelian sampling, unlike pedigree-based relationships. It is important to optimize haploblock length because if haplotype alleles are too short they will not have an improved ability over SNPs to capture IBD relationships, while if they are too long they will underestimate the relationships between close relatives (Ferdosi et al., 2016).

Haplotype alleles may be in higher LD with linked QTL than the high-MAF SNPs that are selected for SNP panels (Zondervan and Cardon, 2004); therefore genomic prediction models that fit covariates for haplotype alleles rather than SNPs may have an improved ability to capture QTL effects. Sun (2014) showed that genomic prediction models that fit haplotype alleles have an improved ability to capture the effects of low-MAF QTL because high-MAF SNPs are unable to be in high LD with low-MAF QTL. Haplotype alleles can be in low frequency, even when generated from high-MAF SNPs, and therefore be in higher LD with these low-MAF QTL. In an admixed-breed population, fitting haplotype alleles rather than SNPs may improve genomic prediction accuracy if haplotype alleles are in higher LD than SNPs with breed-specific QTL. If haploblocks are the appropriate length, and SNP density is adequate, haplotype alleles will be specific to breed and therefore QTL that are segregating in only one breed will be able to be captured in that breed and haplotype alleles specific to another breed where that QTL is not segregating can be assigned estimates of zero.

Haplotype alleles may also be able to capture short-range epistatic effects, e.g. interactions between adjacent genes or between variants within a gene and variants within upstream regulatory elements. The major-histocompatibility-complex (MHC) is a very gene-dense region of the genome and proteins from MHC genes interact to protect the individual from foreign objects (e.g. viruses; Traherne, 2008). Fitting haplotype alleles in genomic prediction models may capture these epistatic effects which may improve prediction accuracy. Fitting haplotype alleles in genomic prediction models may also capture the interactions of variants within genes or between a variant within a gene and its up- or down-stream regulatory elements (Littlejohn et al, 2014; Kuehn et al., 2004).

Impact of Genomic Prediction Method

It is important to compare prediction accuracy for a variety of genomic prediction models when fitting haplotype alleles, as it is for the SNP model. Cuyabano et al. (2015a) found improved prediction accuracy when fitting haplotype alleles rather than SNPs when fitting a Bayesian mixture model (e.g. BayesB) but not when fitting a Bayesian GBLUP model. Ferdosi et al. (2016) recently showed that it is possible to improve prediction accuracy when fitting haplotype alleles in a GBLUP model; however, this model was found to be sensitive to haploblock length and different traits had different optimal haploblock length. BayesN (Zeng, 2015) may be a suitable model for use in genomic prediction when fitting haplotypes when each haploblock is a different window because rather than assuming all haplotype alleles are independent, it will collectively sample a haploblock or not; then sample effects for haplotype alleles within that haploblock if it is sampled in an iteration.

Limitations of Haplotype Models

Although genomic prediction models have shown promise in improving accuracy, there are also some drawbacks. At a density of ~50,000 SNPs haplotype models tend to fit more covariates than SNP models, therefore they take longer to run and require more memory. Approaches to decrease the number of haplotype alleles fitted in genomic prediction models have included only fitting haplotypes in regions with known or putative QTL (Boichard et al., 2012) or removing SNPs with a low minor allele frequency prior to generating the haplotype alleles (Calus et al., 2009; Cuyabano et al., 2015b).

1.5 Evaluating Performance of Genomic Prediction Models

Prediction Accuracy and Bias

The performance of genomic prediction models can be evaluated by splitting the data set into training and validation sets of animals. The training set is used to obtain marker effect estimates and model performance is evaluated via prediction accuracy and bias in the validation set (Daetwyler et al., 2013). Prediction accuracy is the standard measurement to assess model performance (Hayes et al., 2009b; Wolc et al., 2011a) because of its direct relationship to genetic gain (Eq. 1.1). Prediction accuracy is commonly calculated as the Pearson Correlation:

$$cor(y_{adj}, \hat{u}) = \frac{cov(y_{adj}, \hat{u})}{\sqrt{var(y_{adj})var(\hat{u})}} \quad [1.9]$$

where y_{adj} is the phenotype (y) adjusted for the non-genetic effects. This correlation can also be divided by the square root of heritability to approximate the correlation between true breeding values (u) and estimated breeding values ($\hat{u}$):

$$\frac{cor(y,\hat{u})}{\sqrt{h^2}} = \frac{cov(y,\hat{u})}{\sqrt{var(y)var(\hat{u})}}\sqrt{\frac{var(y)}{var(u)}}$$

$$= \frac{cov(y,\hat{u})}{\sqrt{var(u)var(\hat{u})}}$$

Assuming $cov(e,\hat{u}) = 0$, $cov(y,\hat{u}) = cov(u,\hat{u})$

$$\frac{cor(y,\hat{u})}{\sqrt{h^2}} = \frac{cov(u,\hat{u})}{\sqrt{var(u)var(\hat{u})}} = cor(u,\hat{u})$$

where y is the (adjusted) phenotype, u is the additive genetic portion of the phenotype (i.e. true breeding value), $\hat{u}$ is the EBV, and $e = y - u$. It is necessary to approximate this correlation because true breeding values are not directly observed; however this approximation can result in correlations above 1 if the estimate of heritability (h^2) is different from the heritability in the validation set.

Prediction bias is estimated as the regression coefficient from regressing y_{adj} on $\hat{u}$ and can be represented as a deviation from 1, the desired slope of the regression coefficient because a regression coefficient of 1 means that a 1 unit increase in the EBV corresponds to a 1 unit increased in the true BV. A regression coefficient >1 will underestimate the true genetic merit of the best individuals, while a regression coefficient <1 will overestimate the true genetic merit of these individuals; therefore, evaluation of bias will indicate whether the model is appropriately capturing variation in BV (Daetwyler et al., 2013). Bias can also affect genetic gain, particularly when pedigree and genomic information are combined into a single breeding value or when animals with the highest EBV are mated more often than those with lower EBV (Daetwyler et al., 2013).

Training and Validation Sets

The goal of evaluating the performance of a genomic prediction model is to estimate how accurate prediction would be in another set of individuals that do not have phenotypes (i.e. the selection candidates). The choice of training and validation data sets can have a large impact on estimates of prediction accuracy and bias. Prediction accuracy has been shown to increase as the training data set increases; however, given a finite data set to test the model, the larger the training set, the smaller the validation set and a validation set that is too small will not give reliable accuracy estimates (Erbe, 2013). The relationships between individuals in the training and validation set should approximate the relationships between the full data set of genotyped individuals and the selection candidates because prediction accuracy relies on these relationships (i.e. prediction accuracy will be higher if more relatives are included in both the training and validation set; Habier et al., 2007). It may be advantageous to combine closely-related breeds into one training set, particularly when there are few individuals in a single-breed training set (Brondum et al., 2011; Pryce et al., 2011; Kachman et al., 2013); however prediction accuracy may differ between breeds, so it is critical to evaluate prediction accuracy separately for each breed.

Reliability of Accuracy Estimates

Obtaining a measurement of how reliable the accuracy estimate is enables hypothesis testing, e.g. whether one model is better than another model, or whether the prediction accuracy of a model is above a certain threshold. Two commonly used methods for evaluating the reliability of an estimate of accuracy are cross-validation and bootstrapping.

A common approach to cross-validation was described by Saatchi et al. (2011) and is based on k-means clustering, whereby genotyped and phenotyped individuals are separated into k groups to reduce relatedness between groups and maximize relatedness within groups. The training set is then specified as k-1 groups, the validation set is the remaining group, and the genomic prediction accuracy estimate is obtained. The training and validation is then repeated k times until an accuracy estimate is obtained for all groups. The mean and standard deviation of prediction accuracy can be calculated across all groups (Daetwyler et al., 2013). Cross-validation is able to obtain a standard error for the accuracy estimate; however this estimate can be sensitive to outlier groups and the relationships between individuals in training and validation sets is not likely to reflect the relationship between the full data set and selection candidates when the data set is clustered into groups based on relatedness. The accuracy of cross-validation also depends on the value of k, i.e. the number of groups the data set it partitioned into (Erbe, 2013); smaller values of k may be desirable because the genomic prediction model needs to be run k times (once when each group is the validation set).

Bootstrapping genomic prediction data sets involves sampling the individuals in either the training or validation sets with replacement until a new set, the size of the original one, is obtained (Pryce et al., 2014). Prediction accuracy is then estimated with the new set and this process is repeated multiple times, often 5,000 - 10,000 times, to get many samples. The estimate of prediction accuracy is then the mean accuracy across all samples and the standard error of this estimate is the standard deviation of the accuracy across all samples. Bootstrapping can be performed in either the training or validation set, however it is more commonly performed in the validation set due to time constraints of running the prediction model multiple times. When bootstrapping is performed on the validation set, it captures variation in accuracy estimates due

to a different set of validation individuals used but does not capture variation due to the uncertainty in model estimates because it is based on running a single genomic prediction model.

1.6 New Zealand Dairy Cattle

New Zealand dairy farming is considered low-input compared to other dairy farming systems in the developed world. Cattle live outdoors year-round and are able to maintain seasonal milk production on a primarily pasture-based diet because grass grows year-round due to New Zealand's temperate climate. This is in contrast to the high-input grain-based diet that is common in North America and Europe. Holstein cows in New Zealand and North America have been selected in these two very different production systems and North American Holsteins do not perform as well as New Zealand Holsteins in the low-input New Zealand production setting (Konig et al., 2005; Macdonald et al., 2007). As such, there is little flow of genetics from North America and Europe to New Zealand compared to the flow of genetics between North America and Europe (de Roos et al., 2008; de Haas et al., 2015).

Population Structure

The New Zealand dairy cattle population primarily consists of two breeds: Holstein Friesian and Jersey; with a small proportion of Ayrshire and other breeds (LIC and DairyNZ, 2015). The majority of dairy cattle in New Zealand are admixed compared to the norm in other countries (Harris, 2005): an individual only needs to have $7/8^{th}$ Holstein Friesian (Jersey) parentage (i.e. based on pedigree) to be considered a Holstein Friesian (Jersey); and only ~25% of inseminations in New Zealand are to a female of the same breed (LIC and DairyNZ, 2015). The admixed descendants of Holstein Friesian and Jersey cattle, referred to as KiwiCross, have

improved milk production, reproduction and health overall compared to the average of their parent-breeds and are an appealing option for New Zealand dairy farmers (LIC and DairyNZ, 2015). KiwiCross semen is marketed by Livestock Improvement Corporation (LIC), a major breeding company in New Zealand. In addition to the recent crossing of New Zealand Holstein Friesian and Jersey cattle, these two breeds have a common ancestry from Dutch and British Friesians that were imported into New Zealand at the end of the 19th and beginning of the 20th century (Jasiorowski et al., 1988; de Roos et al., 2008). An analysis of low-density SNPs (~9000) found that haplotype phase was highly conserved between Holstein Friesians and Jerseys in New Zealand, almost as high as between New Zealand and Australian Holsteins (de Roos et al., 2008). Genetic evaluations are carried out on the New Zealand dairy cattle population as a whole, rather than separate evaluations by breed, due to the high level of crossbreeding and common ancestry of the major New Zealand dairy cattle breeds (DairyNZ, 2014).

Genomic Selection

LIC was quick to adopt genomic selection, and encouraged the use of genomically evaluated young bulls (Wiggans et al., 2011). LIC initially blended genomic breeding values with the available pedigree information using selection index theory (Harris and Johnson, 2010) to generate high-reliability GEBVs (Spelman et al., 2010). Since then, Harris et al. (2012) have developed single-step evaluation methods to generate these GEBVs. Early genomic prediction in New Zealand suffered from inflated breeding value estimates of top bulls, which is problematic when comparing EBV of genotyped bulls with non-genotyped bulls, but adjustments have been made to overcome this bias (Spelman et al., 2013). A large sequencing project has been initiated by LIC that aims to identify causal variants that may lead to further improvements in genomic

prediction through increased accuracy or reduced bias (Spelman et al., 2010). As mentioned above, genomic evaluations are performed using all breeds, so improved modelling of breed-specific effects by fitting haplotype alleles rather than SNPs may improve genomic prediction accuracy in this population.

1.7 Conclusions

Genomic selection has been successfully implemented in many livestock species because of the greater accuracy of selection in young animals (Hill, 2014). Genomic prediction models that fit haplotype alleles rather than SNPs have the potential to improve accuracy of genomic prediction if they capture the effects of QTL better than SNPs. This may be particularly relevant for the admixed New Zealand dairy cattle population if haplotype alleles are able to capture breed-specific QTL. Haploblocks that account for recombination events and patterns of LD may improve genomic prediction accuracy compared to haploblocks that are the same, arbitrary length across the genome. Obtaining a reliable estimate of the reliability an accuracy estimate, i.e. through a standard error of the estimate, is important for objective model comparison.

1.8 Research Objectives

The overall objective of this dissertation was to investigate whether prediction accuracy, and thus rate of genetic gain, can be improved in the admixed New Zealand dairy cattle population by fitting covariates for haplotype alleles rather than covariates for SNPs in genomic prediction models. Chapter II evaluates the prediction accuracy and bias obtained from fitting covariates for fixed-length haplotypes across the genome, ranging in size from 125 kb to 2 Mb. Chapter III compares the performance of genomic prediction fitting variable-length haplotype

alleles to fixed-length haplotype alleles or SNPs. Chapter IV describes a method for evaluating the uncertainty in Bayesian genomic prediction accuracy estimates through the posterior distribution of prediction accuracy. Chapter V contains a general discussion and identifies further areas of research to improve our ability to make rapid genetic gains in livestock populations.

1.9 References

Arias, J. A., M. Keehan, P. Fisher, W. Coppieters, and R. Spelman. 2009. A high density linkage map of the bovine genome. Bmc Genetics 10doi: 10.1186/1471-2156-10-18

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-265. doi: 10.1093/bioinformatics/bth457

Baudat, F., and A. Nicolas. 1997. Clustering of meiotic double-strand breaks on yeast chromosome III. Proceedings of the National Academy of Sciences of the United States of America 94(10):5213-5218. doi: 10.1073/pnas.94.10.5213

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Animal Production Science 52(2-3):115-120. (Review) doi: 10.1071/an11119

Borde, V., N. Robine, W. Lin, S. Bonfils, V. Geli, and A. Nicolas. 2009. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. Embo Journal 28(2):99-111. doi: 10.1038/emboj.2008.257

Brondum, R. F., E. Rius-Vilarrasa, I. Stranden, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. Journal of Dairy Science 94(9):4700-4707. (Article) doi: 10.3168/jds.2010-3765

Browning, B. L., and S. R. Browning. 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. American Journal of Human Genetics 84(2):210-223. doi: 10.1016/j.ajhg.2009.01.005

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81(5):1084-1097. doi: 10.1086/521987

Browning, S. R., and B. L. Browning. 2011. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics 12(10):703-714. doi: 10.1038/nrg3054

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178(1):553-561. (Article) doi: 10.1534/genetics.107.080838

Calus, M. P. L., T. H. E. Meuwissen, J. J. Windig, E. F. Knol, C. Schrooten, A. L. J. Vereijken, and R. F. Veerkamp. 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. Genetics, Selection, Evolution 41(11):(15 January 2009). (article)

Crawford, D. C., T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder, D. A. Nickerson, and M. Stephens. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nature Genetics 36(7):700-706. doi: 10.1038/ng1376

Cuyabano, B. C. D., G. Su, G. J. M. Rosa, M. S. Lund, and D. Gianola. 2015a. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. Journal of Dairy Science 98(10):7351-7363. doi: 10.3168/jds.2015-9360

Cuyabano, B. C. D., G. S. Su, and M. S. Lund. 2015b. Selection of haplotype variables from a high-density marker map for genomic prediction. Genetics Selection Evolution 47:11. (Article) doi: 10.1186/s12711-015-0143-3

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193(2):347-+. doi: 10.1534/genetics.112.147983

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185(3):1021-1031. doi: 10.1534/genetics.110.116855

Daetwyler, H. D., A. A. Swan, J. H. J. van der Werf, and B. J. Hayes. 2012. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genetics Selection Evolution 44doi: 10.1186/1297-9686-44-33

DairyNZ. 2014. Animal Evaluation. http://www.dairynz.co.nz/animal/animal-evaluation/ (Accessed June 6 2014).

de Haas, Y., J. E. Pryce, M. P. L. Calus, E. Wall, D. P. Berry, P. Lovendahl, N. Krattenmacher, F. Miglior, K. Weigel, D. Spurlock, K. A. Macdonald, B. Hulsegge, and R. F. Veerkamp. 2015. Genomic prediction of dry matter intake in dairy cattle from an international data set consisting of research herds in Europe, North America, and Australasia. Journal of Dairy Science 98(9):6522-6534. doi: 10.3168/jds.2014-9257

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193(2):327-+. doi: 10.1534/genetics.112.143313

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179(3):1503-1512. (Article) doi: 10.1534/genetics.107.084301

Dekkers, J. C. M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. Journal of Animal Breeding and Genetics 124(6):331-341.

Druet, T., and M. Georges. 2010. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184(3):779-798. (Article)

Druet, T., and M. Georges. 2015. LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. Bioinformatics 31(10):1677-1679. doi: 10.1093/bioinformatics/btu859

Erbe, M. 2013. Assessment of cross-validation strategies for genomic prediction in cattle, *Accuracy of Genomic Prediction in Dairy Cattle*.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2014. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 97(10):6622-6622. doi: 10.3168/jds.2014-97-10-6622

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. Introduction to quantitative genetics. (Ed. 4):xv + 464 pp. (Book)

Ferdosi, M. H., J. Henshall, and B. Tier. 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genetics Selection Evolution

Fernando, R., and D. Garrick. 2013. Bayesian Methods Applied to GWAS, Genome-Wide Association Studies and Genomic Prediction. Springer. p. 237-274.

Fernando, R. L., J. C. M. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution 46doi: 10.1186/1297-9686-46-50

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43doi: 10.1186/1297-9686-43-1

Gianola, D. 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194(3):573-596. (Article) doi: 10.1534/genetics.113.151753

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183(1):347-363. doi: 10.1534/genetics.109.103952

Gondro, C., J. Van der Werf, B. J. Hayes, and eds. 2013. Genome-wide Association Studies and Genomic Prediction. Humana Press.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389-2397. (Article) doi: 10.1534/genetics.107.081190

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. Bmc Bioinformatics 12:12. (Article) doi: 10.1186/1471-2105-12-186

Harris, B. L. 2005. Breeding dairy cows for the future in New Zealand. New Zealand Veterinary Journal 53(6):384-389. (Review) doi: 10.1080/00480169.2005.36582

Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. Journal of Dairy Science 93(3):1243-1252. doi: 10.3168/jds.2009-2619

Harris, B. L., A. M. Winkelman, and D. L. Johnson. 2012. Large-scale single-step genomic evaluation for milk production traits No. 46. Interbull Bulletin.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41doi: 5110.1186/1297-9686-41-51

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92(2):433-443. (Review) doi: 10.3168/jds.2008-1646

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetics Research 89(4):215-220. (Article) doi: 10.1017/s0016672307008865

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. Plos Genetics 6(9)doi: 10.1371/journal.pgen.1001139

Heidaritabar, M., M. P. L. Calus, H. J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics

Henderson, C. R. 1984. Applications of linear models in animal breeding. Applications of linear models in animal breeding.:xxiii + 462 pp. (Book)

Hickey, J. M., B. P. Kinghorn, B. Tier, S. A. Clark, J. H. J. van der Werf, and G. Gorjanc. 2013. Genomic evaluations using similarity between haplotypes. Journal of Animal Breeding and Genetics 130(4):259-269. doi: 10.1111/jbg.12020

Hill, W. G. 2014. Applications of Population Genetics to Animal Breeding, from Wright, Fisher and Lush to Genomic Prediction. Genetics 196(1):1-16. doi: 10.1534/genetics.112.147850

Jasiorowski, H. A., M. Stolzman, and Z. Reklewski. 1988. The international Friesian strain comparison trial: A world perspective. Food and agriculture organization of the United Nations.

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers, and P. Donnelly. 2005. Human recombination hot spots hidden in regions of strong marker association. Nature Genetics 37(6):601-606. doi: 10.1038/ng1565

Kachman, S. D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn, W. M. Snelling, R. M. Thallman, M. Saatchi, D. J. Garrick, R. D. Schnabel, J. F. Taylor, and E. J. Pollak. 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genetics Selection Evolution 45:9. (Article) doi: 10.1186/1297-9686-45-30

Kauppi, L., A. J. Jeffreys, and S. Keeney. 2004. Where the crossovers are: Recombination distributions in mammals. Nature Reviews Genetics 5(6):413-424. doi: 10.1038/nrg1346

Kuehn, C., G. Thaller, A. Winter, O. R. P. Bininda-Emonds, B. Kaupe, G. Erhardt, J. Bennewitz, M. Schwerin, and R. Fries. 2004. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative tip trait locus with major effect on milk fat content in cattle. Genetics 167(4):1873-1881.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. Journal of Animal Science 88(2):544-551. (Article) doi: 10.2527/jas.2009-2064

Konig, S., G. Dietl, I. Raeder, and H. H. Swalve. 2005. Genetic relationships for dairy performance between large-scale and small-scale farm conditions. Journal of Dairy Science 88(11):4087-4096.

LIC, and DairyNZ. 2015. New Zealand Dairy Statistics 2014-15, http://www.dairynz.co.nz/media/3136117/new-zealand-dairy-statistics-2014-15.pdf.

Littlejohn, M. D., K. Tiplady, T. Lopdell, T. A. Law, A. Scott, C. Harland, R. Sherlock, K. Henty, V. Obolonkin, K. Lehnert, A. MacGibbon, R. J. Spelman, S. R. Davis, and R. G. Snell. 2014. Expression Variants of the Lipogenic AGPAT6 Gene Affect Diverse Milk Composition Phenotypes in Bos taurus. Plos One 9(1):12. (Article) doi: 10.1371/journal.pone.0085757

Lush, J. L. 1937. Animal breeding plans. Animal breeding plans.:Pp. x + 350. (Book)

Macdonald, K. A., L. R. McNaughton, G. A. Verkerk, J. W. Penno, L. J. Burton, D. P. Berry, P. J. S. Gore, J. A. S. Lancaster, and C. W. Holmes. 2007. A comparison of three strains of

Holstein-Friesian cows grazed on pasture: Growth, development, and puberty. Journal of Dairy Science 90(8):3993-4003. (Article) doi: 10.3168/jds.2007-0119

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. Bmc Genomics 17doi: 10.1186/s12864-016-2443-6

Marston, A. L., and A. Amon. 2004. Meiosis: Cell-cycle controls shuffle and deal. Nature Reviews Molecular Cell Biology 5(12):983-997. doi: 10.1038/nrm1526

Meuwissen, T., and M. Goddard. 2010. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics 185(2):623-U338. (Article) doi: 10.1534/genetics.110.116590

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating Improvement of Livestock with Genomic Selection. In: H. A. Lewin and R. M. Roberts, editors, Annual Review of Animal Biosciences, Vol 1. Annual Review of Animal Biosciences No. 1. Annual Reviews, Palo Alto. p. 221-237.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829. (Article)

Mezard, C. 2006. Meiotic recombination hotspots in plants. Biochemical Society Transactions 34:531-534.

Mourad, R., C. Sinoquet, C. Dina, and P. Leray. 2011. Visualization of Pairwise and Multilocus Linkage Disequilibrium Structure Using Latent Forests. Plos One 6(12)doi: 10.1371/journal.pone.0027320

Mulder, H. A. 2016. Genomic Selection Improves Response to Selection in Resilience by Exploiting Genotype by Environment Interactions. Frontiers in Genetics 7

Myers, S. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321.

Nejati-Javaremi, A., C. Smith, and J. P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75(7):1738-1745.

O'Connell, J., D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, and J. Marchini. 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. Plos Genetics 10(4)doi: 10.1371/journal.pgen.1004234

Paigen, K., and P. Petkov. 2010. Mammalian recombination hot spots: properties, control and evolution. Nature Reviews Genetics 11(3):221-233. doi: 10.1038/nrg2712

Pryce, J. E., O. Gonzalez-Recio, J. B. Thornhill, L. C. Marett, W. J. Wales, M. P. Coffey, Y. de Haas, R. F. Veerkamp, and B. J. Hayes. 2014. Short communication: Validation of genomic breeding value predictions for feed intake and feed efficiency traits. Journal of Dairy Science 97(1):537-542. doi: 10.3168/jds.2013-7376

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Solkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. Journal of Dairy Science 94(5):2625-2630. doi: 10.3168/jds.2010-3719

Rendel, J. M., and A. Robertson. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. Journal of Genetics 50(1):1-8. (Article) doi: 10.1007/bf02986789

Rinaldo, A., S. A. Bacanu, B. Devlin, V. Sonpar, L. Wasserman, and K. Roeder. 2005. Characterization of multilocus linkage disequilibrium. Genetic Epidemiology 28(3):193-206. doi: 10.1002/gepi.20056

Robertson, K. D., and A. P. Wolffe. 2000. DNA methylation in health and disease. Nature Reviews Genetics 1(1):11-19. doi: 10.1038/35049533

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. M. Dekkers, R. L. Fernando, R. D. Schnabel, D. J. Garrick, and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics Selection Evolution 43doi: 10.1186/1297-9686-43-40

Saatchi, M., R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. Bmc Genomics 15:16. (Article) doi: 10.1186/1471-2164-15-442

Sandor, C., W. B. Li, W. Coppieters, T. Druet, C. Charlier, and M. Georges. 2012. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. Plos Genetics 8(7):13. (Article) doi: 10.1371/journal.pgen.1002854

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123(4):218-223. (Article) doi: 10.1111/j.1439-0388.2006.00595.x

Shen, X., M. Alam, F. Fikse, and L. Ronnegard. 2013. A Novel Generalized Ridge Regression Method for Quantitative Genetics. Genetics 193(4):1255-1268. doi: 10.1534/genetics.112.146720

Spelman, R., J. Arias, M. Keehan, V. Obolonkin, A. Winkelman, D. Johnson, and B. Harris. 2010. Application of genomic selection in the New Zealand dairy cattle industry. In: 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany

Spelman, R. J., B. J. Hayes, and D. P. Berry. 2013. Use of molecular technologies for the advancement of animal breeding: genomic selection in dairy cattle populations in Australia, Ireland and New Zealand. Animal Production Science 53(9):869-875. doi: 10.1071/an12304

Stranden, I., and D. J. Garrick. 2009. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of Dairy Science 92(6):2971-2975. doi: 10.3168/jds.2008-1929

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. R. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (similar to 54,000) and high-density (similar to 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95(8):4657-4665. (Article) doi: 10.3168/jds.2012-5379

Sun, X., H. Su, P. Boddhireddy, and D. Garrick. 2016. Haplotype-based Genomic Prediction of Breeds Not in Training. In: Plant and Animal Genomes Conference XXIV, San Diego, CA

Traherne, J. A. 2008. Human MHC architecture and evolution: implications for disease association studies. International Journal of Immunogenetics 35(3):179-192. doi: 10.1111/j.1744-313X.2008.00765

Tuiskula-Haavisto, M., D. J. De Koning, M. Honkatukia, N. F. Schulman, A. Maki-Tanila, and J. Vilkki. 2004. Quantitative trait loci with parent-of-origin effects in chicken. Genetical Research 84(1):57-66. doi: 10.1017/s0016672304006950

van Binsbergen, R., M. P. L. Calus, M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47:13. (Article) doi: 10.1186/s12711-015-0149-x

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91(11):4414-4423. (Article) doi: 10.3168/jds.2007-0980

Villumsen, T. M., and L. Janss. 2009. Bayesian genomic selection: the effect of haplotype length and priors. BMC proceedings 3 Suppl 1:S11.

Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics 126(1):3-13. (Article) doi: 10.1111/j.1439-0388.2008.00747.x

Weng, Z. Q., M. Saatchi, R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genetics Selection Evolution 46doi: 10.1186/1297-9686-46-34

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. Journal of Dairy Science 94(6):3202-3211. doi: 10.3168/jds.2010-3866

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011a. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genetics Selection Evolution 43doi: 10.1186/1297-9686-43-23

Wolc, A., C. Stricker, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, S. J. Lamont, and J. C. M. Dekkers. 2011b. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genetics Selection Evolution 43doi: 10.1186/1297-9686-43-5

Wright, S. 1922. Coefficients of inbreeding and relationship. American Naturalist 56:330-338. doi: 10.1086/279872

Zeng, J. 2015. Whole genome analyses accounting for structures in genotype data, Iowa State University, http://lib.dr.iastate.edu/etd/14699.

Zondervan, K. T., and L. R. Cardon. 2004. The complex interplay among factors that influence allelic association. Nature Reviews Genetics 5(2):89-U14. (Review) doi: 10.1038/nrg1270

CHAPTER II

FIXED-LENGTH HAPLOTYPES CAN IMPROVE GENOMIC PREDICTION ACCURACY IN AN ADMIXED DAIRY CATTLE POPULATION

Melanie Hayr[1,2], Tom Druet[3], Andrew Hess[1], and Dorian Garrick[1,4]

[1]Iowa State University, Iowa, USA

[2]LIC, Hamilton, New Zealand

[3]University of Liege, Liege, Belgium

[4]Massey University, Palmerston North, New Zealand

A paper submitted to *Genetics Selection Evolution*

2.1 Abstract

Background

Fitting covariates that represent the number of haplotype alleles rather than SNP alleles may increase genomic prediction accuracy if there is inadequate linkage disequilibrium between QTL and SNPs. Approximately 58,000 New Zealand dairy cattle were genotyped on Illumina BovineSNP50 or HD panels. Genotypes at 37,740 SNPs were phased. All genotyped females born before 1 June 2008 were used for training, and genomic predictions for milk fat yield (n = 24,823), liveweight (n = 13,283) or somatic cell score (n = 24,864) were validated in later-born genotyped females within breed (predominantly Holstein Friesian, predominantly Jersey, or

admixed KiwiCross). Haplotype alleles were determined based on phased SNP alleles within non-overlapping blocks of 125 kb, 250 kb, 500 kb, 1 Mb or 2 Mb. Haplotype alleles with frequency below 1, 2.5, 5 or 10% were removed. Genomic predictions fitting covariates for either SNPs or haplotypes used BayesA, BayesB or the regionally hierarchical model, BayesN.

Results

The correlation between genomic predictions obtained using BayesA and the yield deviation for milk fat in Holstein Friesians was 0.348±0.016 when using SNPs. Using shorter haplotypes resulted in more accurate predictions (0.356±0.016; $p = 0.003$; for lengths of 250 kb with 1% frequency), whereas using longer haplotypes was less accurate than SNPs. Similar trends were apparent for all traits and breeds. For shorter haplotypes, fitting only more frequent haplotype alleles reduced computing effort with little decline in prediction accuracy. There was little practical difference between Bayesian methods.

Conclusions

Fitting covariates for haplotype alleles rather than SNPs slightly increases genomic prediction accuracy when haplotype lengths are appropriately defined. Fitting haplotypes of length 250 kb with a 1% frequency threshold resulted in the highest accuracy of genomic prediction. Accuracies were similar for the different Bayesian methods used, but BayesB models required much less computing time than BayesA or BayesN models. These improvements in accuracy are likely to increase genetic gain by changing the ranking of selection candidates.

2.2 Background

Availability of SNP genotypes has enabled the estimation of breeding values with higher accuracy at a young age than breeding values based on parent average (Van Raden, 2008).

Genomic prediction is routinely performed by fitting covariates representing SNP allele dosage, putatively relying on linkage disequilibrium (LD) between SNPs and QTL to estimate the QTL effects (Habier et al., 2007; Meuwissen et al., 2013). A haplotype block (haploblock) defines a region of the genome that comprises a set of neighboring genetic markers (i.e. SNPs) whose phased alleles are likely to be inherited together. A haplotype allele is a combination of phased SNP alleles that are present in a haploblock. Haplotype alleles are likely in higher LD with a linked QTL than the high-minor allele frequency (MAF) non-coding SNP alleles that are typically used to populate SNP chips (Zondervan and Cardon, 2004). If the LD between haplotype alleles and QTL within the haploblock is higher than between individual SNP alleles and QTL, the accuracy of genomic predictions that fit covariates for haplotype alleles rather than SNP alleles is expected to be higher.

The predictive accuracy of haplotype models has been shown to be influenced by the method used to divide the genome into haploblocks, in both simulated (Villumsen and Janss, 2009; Villumsen et al., 2009) and real (Hayes et al., 2007) data. Simple methods to form haploblocks use measurements of length, such as centimorgans (Boichard et al., 2012), base pairs (Sun et al., 2016) or the number of SNPs (Hayes et al., 2007; Calus et al., 2009; Villumsen et al., 2009), and apply these uniformly along the genome. These fixed-length haplotypes are easy to construct and their definition is not sensitive to the dataset used to construct them, unlike more complex methods (Calus et al., 2008; Cuyabano et al., 2015b) that attempt to account for recombination hotspots and coldspots along the genome (Sandor et al., 2012; Weng et al., 2014).

Discarding SNPs with low MAF is common practice when performing genomic prediction in order to reduce computation time and because there is little power to detect trait associations for SNPs with low MAF (VanRaden et al., 2009; Turner et al., 2011). There are

over 1 million possible haplotype alleles for a block of 20 SNPs, and although far fewer are found in practice, many are typically observed at low frequency. Discarding these rare haplotype alleles will reduce computation time with little expected decrease in prediction accuracy, because the effects of rare alleles is shrunk towards zero in Bayesian linear regression models (Gianola, 2013).

Cuyabano et al. (2015a) found that fitting covariates for haplotype alleles instead of SNPs increased the accuracy of genomic predictions when fitting a Bayesian mixture model but not when fitting a RR-BLUP model. BayesA (Meuwissen et al., 2001) fits all markers simultaneously and marker effects are assumed to be independent with a marker-specific variance. Not all genomic regions are expected to be associated with the phenotype. BayesB (Meuwissen et al., 2001) defines a parameter π and samples marker effects from mixture distributions, whereby the effects for approximately 1-π of markers are sampled in each iteration of a Markov chain with the same assumptions as BayesA, and the remainder are assumed to be zero. BayesN (Zeng, 2015) is a hierarchical extension to BayesB that assumes some chromosome segments have non-zero effects and applies a local BayesB model to only those chromosome segments that are sampled to have an effect. Its hyperparameters include Π, the proportion of segments that are assumed to have no effect, which dictates that a proportion of approximately 1-Π of segments is sampled to have non-zero effects in each iteration; and πi, the segment-specific probability that a covariate within that segment has a zero effect. We hypothesized that BayesN would perform well when fitting covariates for haplotype alleles if each haploblock is considered a window because it will estimate non-zero effects for those haplotype alleles that are in genomic regions (haploblocks) associated with the phenotype, and effects of zero for covariates in all other regions.

Most studies using haplotypes to improve genomic prediction accuracy have focused on simulated datasets (Calus et al., 2008; Villumsen et al., 2009; Hickey et al., 2013), or datasets consisting of a single breed (Hayes et al., 2007; de Roos et al., 2011; Cuyabano et al., 2015b). The New Zealand dairy cattle population consists predominantly of Holstein Friesians, Jerseys, or their admixed descendants, known as KiwiCross (KX). Bulls used for AI include KX in addition to bulls that are predominantly (≥ 7/8) Holstein Friesian (HF) or predominantly (≥ 7/8) Jersey (J); only ~25% of inseminations in New Zealand are typically to a female of the same breed (i.e. HF, J or Ayrshire) (LIC and DairyNZ, 2015). This makes the majority of New Zealand dairy cattle admixed in contrast to the norm in other countries (Harris, 2005). This is the first study to quantify the accuracy of genomic prediction using haplotypes across the whole genome in an admixed population.

The objectives of this study were to evaluate the accuracy, bias and computing time of Bayesian genomic prediction methods that fit covariates for fixed-length haplotype alleles compared to SNP alleles. Fixed-length haplotype alleles, ranging from 125 kb to 2 Mb in length, with varying allele frequency thresholds, ranging from 1 to 10%, were fitted using BayesA (Meuwissen et al., 2001), BayesB (Meuwissen et al., 2001), and BayesN (Zeng, 2015) models for genomic prediction when training using all breeds and validating the resulting predictions in later-born HF, J or KX cows that were not included in training.

2.3 Methods

Phenotype Data

First lactation yield deviations (YD) (Van Raden and Wiggans, 1991) were provided by Livestock Improvement Corporation (LIC) for milk fat yield (Fat), liveweight (Lwt) and somatic

cell score (SCS) for cows that were born between 1990 and 2011. Heritabilities of these traits were assumed to be 0.28, 0.30 and 0.15, respectively (LIC, 2009). Records were discarded if the animal was >1/16 of any breed other than Holstein, Friesian, Jersey or Red Dairy Cattle (e.g. Ayrshire) according to a six-generation pedigree. All animals in small (<5 records) contemporary groups (same herd, parity, calving season, and test day), outlier contemporary groups and outliers within a contemporary group were excluded. Outliers were defined as those animals or groups whose records deviated >5 sd from the mean for Fat and Lwt or >7 sd for SCS. Genotyped females with YD were used for training if they were born before 1 June 2008, and later-born genotyped females comprised the validation data. June 1st is the recognized start of the New Zealand Spring calving season. The number of animals in each training and validation sets by breed is in Table 2.1.

Table 2.1: Numbers of records in training and validation sets used for genomic prediction

Breed[1]	**Fat**[2]		**Lwt**[2]		**SCS**[2]	
	Training	Validation	Training	Validation	Training	Validation
HF	9072	3354	3908	1464	9094	3358
J	5067	5854	2667	2331	5071	5860
KX	10684	6125	6708	2436	10699	6140
Total	24823[3]	15333	13283[3]	6231	24864[3]	15358

1) HF = predominantly (>7/8) Holstein Friesian; J = predominantly (>7/8) Jersey; KX = admixed KiwiCross
2) Yield Deviation: Fat = Milk Fat Yield; Lwt = Liveweight; SCS = Somatic Cell Score
3) Training was performed using pooled data across the three breed classes

Genotypes and Phasing

Genotype information was collected on either v1 or v2 Illumina BovineSNP50 Beadchips (Matukumalli et al., 2009) or the Illumina BovineHD Beadchip (Matukumalli et al., 2011) for 58,369 dairy cattle born between 1960 and 2012 (females = 46,614; males = 11,755). Filtering based on Hardy Weinberg Equilibrium ($p < 1e\text{-}8$), call rate (<0.95) and excess Mendelian inconsistencies (>10) left 37,802 mapped autosomal SNPs that were phased using

LINKPHASE3 (Druet and Georges, 2015). Markers associated with 35 putative map errors (Druet and Georges, 2015) were removed, leaving 37,740 SNPs. Some regions remained un-phased for some individuals, and these regions were phased with DAGPHASE (Druet and Georges, 2010) using the Directed Acyclic Graph obtained from all haplotypes phased with BEAGLE (Browning and Browning, 2007).

Haplotype Construction

Haplotypes of five different lengths (125 kb, 250 kb, 500 kb, 1 Mb and 2 Mb) were constructed using the UMD 3.1 map of the *Bos taurus* genome (Genbank accession: DAAA00000000.2). Rare haplotype alleles were discarded based on their frequency in the training dataset at four different frequency thresholds: 1, 2.5, 5 or 10%. The five choices of haplotype length assessed at each of the four frequency thresholds resulted in 20 scenarios for each haplotype model.

Genomic Prediction Models

Genomic Prediction was performed using GenSel v4.73R (Garrick and Fernando, 2013), by fitting covariates for either SNPs or haplotype alleles in BayesA, BayesB or BayesN models. A single Markov chain Monte Carlo (MCMC) chain of length 41,000, including burn-in of 1,000 iterations, was computed for each analysis to obtain posterior estimates of covariate effects, which were used to obtain direct genomic values (DGV) for validation animals, as described in the following. Prior analysis had shown that correlations and regression coefficients had converged at this chain length.

BayesA

The SNP model and all 20 scenarios of the haplotype model, comprising 5 haplotype lengths and 4 haplotype allele frequency thresholds were fitted using BayesA for all traits, using the following model (Meuwissen et al., 2001):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Xh} + \sum_{j=1}^{k} \boldsymbol{z}_j \alpha_j + \boldsymbol{e} \qquad [2.1]$$

where **y** is an N × 1 vector of YD, μ is the intercept, **X** is an incidence matrix of pairwise heterosis fractions between Holstein (H), Friesian (F), Jersey (J) and Red (R) breeds, defined as the product of the pedigree-based proportions of each of the two breeds for an individual, **h** is a vector of 6 heterosis effects, k is the number of covariates for SNPs (SNP model) or haplotype alleles (haplotype model), $\mathbf{z}_j$ is an N × 1 vector of allele counts (0/1/2) at SNP j (SNP model) or haplotype allele j (haplotype model), α_j is the additive effect of that SNP or haplotype allele, and **e** is an N × 1 vector of identically and independently distributed residual effects with a mean zero and variance σ_e^2, where the prior for σ_e^2 is a scaled inverse chi-square distribution with scale parameter S_e^2 and ν_e degrees of freedom. BayesA assumes that marker effects have identical and independent t-distributions with scale parameter S_α^2 and ν degrees of freedom.

BayesB

The SNP model and two of the twenty haplotype scenarios were fitted using BayesB. The haplotype scenarios chosen were the most accurate scenario based on BayesA across all breeds and traits and a scenario that fitted a similar number of covariates as the SNP model. The BayesB model (Meuwissen et al., 2001) can be written as:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Xh} + \sum_{j=1}^{k} \boldsymbol{z}_j \alpha_j \delta_j + \boldsymbol{e} \qquad [2.2]$$

where variables are defined as for BayesA, except that BayesB is a mixture model that assumes some of the α_j have zero effect. This is defined by the binary variable δ_j that represents whether covariate j was fitted in the model according to hyperparameter π, such that δ_j =1 with probability 1-π, or $\delta_j = 0$ with probability π. BayesA is identical to BayesB when $\pi = 0$. A range of values of π (0.2, 0.35, 0.5, 0.65, 0.8 and 0.95) were compared for all traits with the SNP and the two haplotype models to evaluate the sensitivity of BayesB to the assumed π.

BayesN

Only the SNP and the two haplotype scenarios that were fitted for BayesB were fitted for BayesN for each trait. The model for BayesN (Zeng et al., 2015) was:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Xh} + \sum_{i=1}^{w}\sum_{j=1}^{m_i} \boldsymbol{z}_{ij}\alpha_{ij}\delta_{ij}\Delta_i + \boldsymbol{e} \qquad [2.3]$$

where variables are defined as for BayesB, except that w is the number of windows (represented by haploblocks for haplotype models) and m_i is the number of covariates (SNPs or haplotype alleles) in window i. Parameter $\mathbf{z}_{ij}$ is an N × 1 vector of allele counts (0/1/2) at SNP j in window i (SNP model) or of haplotype allele j in window i (haplotype model), α_{ij} is the additive effect of that SNP or haplotype allele. Marker effects were assumed to have identical and independent mixture distributions of zero with probability Π and t-distribution with scale parameter S_α^2 and ν degrees of freedom with probability 1-Π. This is a different approach than Zeng et al. (2015), who sampled with a window-specific variance. Parameter Δ_i is a binary variable that represents whether covariates in window i are sampled with the same assumptions as BayesB ($\Delta_i = 1$ with probability 1-Π) or are sampled to have a zero effect ($\Delta_i = 0$ with probability Π). A range of values for Π (0.5, 0.8 or 0.95) were assumed to test the sensitivity of BayesN to Π. The GenSel

implementation of BayesN fitted k covariates per window, therefore $\delta_{ij} = 1$ with probability $1-\pi_i$ and $\delta_{ij} = 0$ with probability π_i where

$$\pi_i = \frac{m_i - k}{m_i} \quad [2.4]$$

and m_i is the number of markers in window i.

Each BayesN SNP model was run twice, once with k = 2, equivalent to fitting BayesB within a sampled window, and once where k was set to the maximum number of SNPs in a window (i.e. $\pi_i = 0$), equivalent to fitting BayesA within a sampled window. Haplotype models were run with $\pi_i = 0$, equivalent to BayesA within a haploblock.

Evaluation of Prediction Models

The training set for all genomic prediction models contained all breed classes (HF, J and KX), but predictions of validation cows were evaluated separately for each breed. The DGV were calculated for validation cows as:

$$\widehat{\boldsymbol{DGV}} = \boldsymbol{X}\hat{\boldsymbol{h}} + \sum_{j=1}^{k} \boldsymbol{z}_j \hat{\alpha}_j \quad [2.5]$$

where heterosis was included because validation used yield deviations which include heterosis. Model performance was evaluated based on prediction accuracy, calculated as the correlation between YD and DGV, and prediction bias, the deviation from 1 of the regression coefficient of YD on DGV.

Bootstrap Samples

Estimation of the accuracy and bias of genomic prediction from the entire validation set does not give an indication of the sampling error associated with the estimate. Therefore,

standard errors were obtained from a single training analysis using 10,000 bootstrap samples of validation animals for each breed. Validation animals within a breed were sampled with replacement to obtain a sample equal in size to the validation set for that breed. The same bootstrap samples of validation animals were used for all scenarios and models. Prediction accuracy and bias were calculated for each bootstrap sample, and the estimate and standard error of these parameters for the validation set were the mean and standard deviation across bootstrap samples. Comparisons between models were obtained from paired t-tests of the 10,000 bootstrap samples, whereby accuracy (or bias) were paired across each model for the same sample of animals. T-tests were one sided when comparing the accuracy of a haplotype model to the accuracy of a SNP model because we were interested in testing whether haplotype models improved prediction accuracy over a SNP model, and two-sided otherwise. Significance was determined based on a p-value threshold of 0.05.

Additional Evaluation Criteria

In addition to accuracy and bias of the models, the number of random effects fitted in the model (SNPs or haplotype alleles) and computation time were evaluated. The mean squared error of the model for the validation set of animals was also assessed.

Potential Impact of Haplotype Models on Selection Decisions

The Spearman rank correlation of DGV from all cows and the top 100 cows were compared between the BayesA SNP and Hap250-1 (250 kb haplotypes, fitting haplotype alleles with frequency >1% in training) models. According to DairyNZ (2009), the top ~0.9% of cows are selected to be dams of the next generation of bulls in New Zealand. Therefore, the number of

cows that were in the top 0.9% for both models was also reported to evaluate whether moving from a SNP model to a haplotype model is likely to impact selection decisions.

2.4 Results

The number of SNPs in each haploblock varied across the genome (Table 2.2). The minimum number of SNPs in a haploblock was 1 for all haplotype lengths. The average number of SNPs per haploblock ranged between 2 and 30 and the maximum ranged from 6 to 54.

Table 2.2: Mean and Maximum Number of SNPs by Haploblock Length

Haploblock Length	**Number of Haploblocks**	**Number of SNPs per Haploblock[1]**	
		Mean	Maximum
125 kb	17452	2	6
250 kb	9676	4	10
500 kb	4978	8	17
1 Mb	2514	15	31
2 Mb	1267	30	54

1) The minimum number of SNPs in a haploblock was 1 for all haplotype lengths

BayesA

Prediction Accuracy and Bias

The prediction accuracy and bias of each BayesA model are in Figures 2.1 (Fat), S2.1 (Lwt) and S2.2, (SCS), and Table S2.3. Liveweight had the highest predictive accuracy of the three traits, followed by Fat then SCS, in accord with their heritabilities. Holstein Friesians had a higher accuracy than Jerseys for all three traits and KiwiCross was intermediate for Fat and SCS but had the highest accuracy for Lwt. The most accurate haplotype models were more accurate than the SNP model and were similarly or less biased. Increasing haploblock length or the

haplotype allele frequency threshold tended to decrease accuracy and increase bias of the haplotype model; however the most accurate model overall utilized haploblocks of 250 kb and a haplotype allele frequency filter of 1% (Figures 2.1, S2.1 and S2.2). Increasing the frequency threshold had a more negative impact on both accuracy and bias for the longer (1 – 2 Mb) haploblocks than shorter (125 – 250 kb) haploblocks. Prediction bias for models that used shorter haploblocks was similar to that for the SNP model, while models that used longer haploblocks were usually more biased than the SNP model (Figures 2.1, S2.1 and S2.2).

Number of Covariates and Computation Time

Table 2.3 contains the number of random covariates fitted in each BayesA model and the computation time in hours for each model, not including the time to generate and filter the haplotype alleles. The number of covariates fitted was similar across the three traits. Fitting 2 Mb haplotype alleles with a frequency threshold of 10% fitted only 650-700 haplotype alleles and was the least accurate model. Computation times increased as the number of haplotype alleles increased. The most accurate model for all three traits took approximately twice as long to run than the SNP model because it fitted approximately twice as many covariates. The fastest models ran in 15 – 20 minutes, depending on the trait (Table 2.3), but this corresponded with a drastic decrease in accuracy and increase in bias (Figures 2.1, S2.1 and S2.2, Table S2.3).

Table 2.3: Computation Time and Number of Random Effects in Haplotype and SNP BayesA Models

Trait	Freq[1]	Number of Random Effects					Computation Time (hours)				
		125 kb	250 kb	500 kb	1 Mb	2 Mb	125 kb	250 kb	500 kb	1 Mb	2 Mb
Milk Fat Yield	SNP	37226	37226	37226	37226	37226	15.0	14.0	13.4	13.1	12.8
	1%	56590	64724	70380	56534	32520	22.8	23.5	24.7	20.0	11.3
	2.5%	51889	53482	47378	29343	13460	21.3	19.7	16.8	10.4	4.8
	5%	46283	41737	28324	12291	3977	19.6	15.5	10.4	4.5	1.5
	10%	37848	27656	12790	3255	646	15.2	10.8	5.0	1.4	0.3
Liveweight	SNP	37356	37356	37356	37356	37356	7.5	7.1	6.8	6.6	6.5
	1%	56595	64634	70218	56164	32117	11.0	13.1	13.3	9.9	5.7
	2.5%	51839	53204	46797	28756	13050	10.2	9.6	9.2	5.2	2.4
	5%	46163	41467	28040	12198	4027	9.2	7.7	5.2	2.3	0.8
	10%	37775	27604	12882	3354	707	7.8	5.4	2.6	0.7	0.2
Somatic Cell Score	SNP	37229	37229	37229	37229	37229	15.0	13.9	13.2	13.0	12.8
	1%	56630	64730	70375	56521	32516	21.4	24.4	27.2	19.7	11.1
	2.5%	51934	53488	47385	29348	13464	23.1	20.8	16.7	10.9	4.7
	5%	46326	41746	28329	12296	3977	18.3	15.4	10.2	4.5	1.5
	10%	37898	27663	12793	3254	645	15.1	10.7	5.0	1.3	0.3

1) Frequency threshold for removing rare haplotype alleles. SNP refers to fitting covariates for SNPs rather than haplotype alleles.

BayesB and BayesN

Haplotype Model Choice

Two scenarios from the BayesA analyses were chosen to evaluate whether a BayesB or a BayesN model would lead to a further increase in accuracy over the BayesA haplotype model: 250 kb haploblocks fitting only alleles present at greater than 1% frequency in the training data set (Hap250-1) and 125 kb haploblocks fitting only alleles present at greater than 10% frequency (Hap125-10). Hap250-1 was selected because the mean square error (MSE) of this model was the lowest for all three traits (Table S2.4); this scenario also had the highest accuracy and consistently low bias (Figures 2.1, S2.1 and S2.2). The Hap125-10 model was selected because there were a similar number of haplotype alleles as there were SNPs (Table 2.3), and could be used to evaluate whether it would be better to fit SNP or haplotype alleles if the number of

covariates has to be constrained. The MSE of the BayesA Hap125-10 model was less than or equal to the MSE of the SNP model for all three traits (Table S2.4).

Prediction Accuracy and Bias

The accuracy of the BayesN SNP model was similar when sampling non-zero effects for all SNPs in a sampled window or only sampling non-zero effects for 2 SNPs in a sampled window (Table S2.5). Window size (125 kb, 250 kb, or 1 Mb) also had very little impact on prediction accuracy for BayesN. Therefore, only results using 250 kb windows and sampling all SNPs per window were further evaluated.

A range of values of π (BayesB) and Π (BayesN), collectively referred to as pi values, were evaluated to determine values that led to the highest accuracies. Accuracies were essentially the same but decreased when pi values were so high that too few features were fitted, corresponding to pi values over 0.8 for most traits and breeds (Figures S2.6 and S2.7): ~7,000 covariates for the SNP and Hap125-10 models and ~12,000 covariates for the Hap250-1 model. In the remainder, BayesB and BayesN results for a pi value of 0.5 will be presented because, in many cases, this value resulted in the highest or close to highest accuracy.

Prediction Accuracy

Bayesian method (i.e. BayesA vs. BayesB vs. BayesN) had very little impact on prediction accuracy for both SNP and haplotype models (Figure 2.2). Haplotype models were more accurate than the SNP model for all traits and breeds, except for Fat in Jerseys, for which prediction accuracy was very similar across all models. Haploblocks of length 250 kb tended to have higher accuracies than haploblocks of length 125 kb but this difference was not significant ($P > 0.24$) except for SCS in Jerseys ($P < 0.006$) and Fat in KiwiCross ($P < 0.077$). Based on these results, the BayesA Hap250-1 model was chosen as a representative model for comparison

with the BayesA SNP model. Compared to the BayesA SNP model, the BayesA Hap250-1 model showed difference in accuracy equal to 2.1 ± 1.1% (Fat; HF); -0.1 ± 1.2% (Fat; J); 2.3 ± 1.0% (Fat; KX); 2.1 ± 1.5% (Lwt; HF); 3.4 ± 1.8% (Lwt; J); 2.2 ± 1.1% (Lwt; KX); -0.1 ± 2.3% (SCS; HF); 5.5 ± 2.1% (SCS; J); and 0.0 ± 2.0% (SCS; KX).

Prediction Bias

Prediction bias was significantly different from zero for all traits in Jerseys, for no trait in Holstein Friesian, and only for SCS in KiwiCross (Table 2.4). Most models did not significantly change prediction bias compared to the BayesA SNP models, and even significant changes were small. Compared to the BayesA SNP model, the BayesN models tended to result in more biased predictions, some of which were significantly different from the BayesA SNP model. Conversely, Bayes A haplotype models and BayesB models resulted in less biased predictions when they were significantly different from the BayesA SNP models.

Number of Covariates and Computation Time

Haplotype models had a longer computing time than SNP models for all analyses, which was driven by the number of covariates that were fitted in each model (Table 2.5). BayesB models had a shorter computing time than the corresponding BayesA model, but BayesN models had a much longer computing time. Computing times for Fat and SCS were approximately double the times for Lwt because the training set had approximately twice the number of records (Table 2.1).

Potential Impact of Haplotype Models on Selection Decisions

The Spearman rank correlation between DGV from the BayesA SNP model and the BayesA Hap250-1 model was high (≥0.95) when considering all cows, but there was a

considerable amount of re-ranking when considering only the top 100 cows for each breed and trait (Table 2.6). This re-ranking had an impact on which cows had DGV in the top 0.9%, suggesting that fitting haplotypes rather than SNPs will impact which animals are selected as dams of sires.

Table 2.4: Prediction bias of SNP and Haplotype Models for BayesA, BayesB and BayesN Analyses

Trait	Breed	**BayesA**			**BayesB (π = 0.5)**			**BayesN (Π = 0.5; π = 0)**		
		SNP	Hap125	Hap250	SNP	Hap125	Hap250	SNP	Hap125	Hap250
Fat	HF	*-0.04*	-0.07	-0.05	-0.05	-0.07	-0.06	-0.01	-0.05	-0.04
		(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
	J	*0.16*	0.16	0.17	**0.16**	0.16	0.16	**0.19**	**0.18**	0.18
		(0.04)	(0.04)	(0.04)	**(0.04)**	(0.04)	(0.04)	**(0.04)**	**(0.04)**	(0.04)
	KX	*-0.01*	-0.02	-0.03	-0.02	-0.02	-0.04	0.02	0.00	-0.02
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Lwt	HF	*0.03*	0.03	0.01	0.03	0.02	0.00	0.07	0.04	0.02
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	J	*0.21*	0.21	**0.18**	**0.20**	0.20	0.18	**0.24**	0.21	0.20
		(0.05)	(0.04)	**(0.05)**	**(0.05)**	(0.04)	(0.05)	**(0.04)**	(0.04)	(0.04)
	KX	*0.00*	0.01	-0.02	-0.01	0.00	-0.03	0.03	0.02	-0.01
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
SCS	HF	*0.05*	0.05	0.05	0.05	0.05	0.05	0.13	0.09	0.09
		(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)
	J	*0.23*	0.22	**0.18**	0.23	0.22	**0.18**	**0.29**	**0.26**	0.22
		(0.07)	(0.07)	**(0.07)**	(0.07)	(0.07)	**(0.07)**	**(0.06)**	**(0.06)**	(0.07)
	KX	*0.17*	0.19	0.16	0.17	0.19	0.16	**0.24**	**0.23**	0.20
		(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	**(0.06)**	**(0.06)**	(0.06)

Bold = Significantly different bias than the BayesA SNP model (*italics)* for that breed and trait ($P < 0.05$)

1) Trait: Fat = Milk Fat Yield; Lwt = Liveweight; SCS = Somatic Cell Score
2) Breed: HF = predominantly Holstein Friesian; J = predominantly Jersey; KX = admixed KiwiCross (HF/J)
3) Hap125 = Haplotypes of length 125 kb, fitting only haplotype alleles >10% frequency in training data set
4) Hap250 = Haplotypes of length 250 kb, fitting only haplotype alleles >1% frequency in training data set

Table 2.5: Number of Random Covariates and Computation Time for Each Model

Model[1]		Number of Random Effects[2]			Computation Time (h)		
		Fat[3]	Lwt[3]	SCS[3]	Fat[3]	Lwt[3]	SCS[3]
BayesA	SNP	37226	37356	37229	13.1	6.6	13.0
	Hap125	37848	37775	37898	15.2	7.8	15.1
	Hap250	64724	64634	64730	23.5	13.1	24.4
BayesB	SNP	18589	18637	18629	10.0	5.1	9.9
	Hap125	18899	18831	18954	13.6	6.2	13.9
	Hap250	32332	32273	32388	18.1	9.2	18.0
BayesN	SNP	17748 (4701)	17639 (4671)	18254 (4805)	26.7	12.5	25.6
	Hap125	18451 (8264)	18303 (8223)	18711 (8344)	30.2	16.0	30.0
	Hap250	31596 (4737)	31281 (4706)	32103 (4809)	37.6	18.9	38.1

1) SNP = SNP model with 250 kb windows
Hap125 = Haplotypes of length 125 kb, fitting only haplotype alleles >10% frequency in training data set
Hap250 = Haplotypes of length 250 kb, fitting only haplotype alleles >1% frequency in training data set
2) Average number of SNPs or haplotype alleles fitted in each chain of the MCMC
3) Fat = Milk Fat Yield; Lwt = Liveweight; SCS = Somatic Cell Score

Table 2.6: Rankings from the BayesA 250kb Haplotype Model Compared to the BayesA SNP Model

Trait	Breed	r_S(All)[1]	r_S (Top 100)[2]	Top 0.9%[3]
Fat	HF	0.97	0.69	23/30
	J	0.97	0.66	40/53
	KX	0.97	0.59	43/55
Lwt	HF	0.97	0.60	10/13
	J	0.95	0.67	12/21
	KX	0.96	0.70	17/22
SCS	HF	0.96	0.63	20/30
	J	0.97	0.67	42/53
	KX	0.96	0.45	34/55

1) Spearman Rank Correlation for all cows
2) Spearman Rank Correlation for the joint set of cows that are in the top 100 cows for DGV from the SNP model or the top 100 cows for DGV from the haplotype model
3) Number of animals with DGV in the top 0.9% for both the SNP model and haplotype model over the number of animals that are in the top 0.9% for the SNP model

2.5 Discussion

Meuwissen and Goddard (2010) predicted a promising increase in genomic prediction accuracy when increasing SNP density from ~30,000 SNPs to sequence. These results have not been observed in practice, with only a slight increase in genomic prediction accuracy has been observed when fitting covariates for SNPs from the Illumina BovineHD panel (~777,000 SNPs) rather than the Bovine SNP50 panel (~54,000 SNPs) (Su et al., 2012; Erbe et al., 2014), and little improvement or even a reduction in accuracy when fitting sequence variants (van Binsbergen et al., 2015; Heidaritabar et al., 2016). Our study highlighted the potential of improving genomic prediction accuracy through the use of haplotypes. Fitting covariates for haplotype alleles rather than SNPs could increase prediction accuracy through improved ability to detect ancestral relationships between individuals (i.e. identity-by-descent), higher LD between causal mutations and haplotype alleles, or greater ability to capture epistasis; likely a mixture of all three. The ability of a haplotype model to improve predictive accuracy depends on the prior assumptions of the model, the method used to define haploblocks and haplotype alleles, SNP density, and the training and validation set demographics.

Haplotype Parameters

Haploblock Length

Villumsen et al. (2009) evaluated optimal haploblock length for simulated traits with heritabilities ranging from 0.02 to 0.30 and found that haploblocks of 1 cM gave the best results across all heritabilities. In New Zealand dairy cattle 1 Mb is equal to approximately 1.25 cM (Arias et al., 2009). Our study also found a single haploblock length that gave the highest prediction accuracy across all the traits investigated, with heritability ranging from 0.15 (SCS) to

0.30 (Lwt), as well as across all breeds (HF, J and KX); however, in our study, much shorter (250 kb) haploblocks had the highest prediction accuracy (Figures 2.1, S2.1 and S2.2).

Prediction accuracies of haplotype models were generally lower than those of the SNP model when haploblocks were greater than 1 Mb in length (>15 SNPs per haploblock on average). This drop in accuracy is likely due to the large number of low frequency haplotype alleles that are generated from long haploblocks, and therefore removed in our analysis. The number of rare haplotype alleles can be observed in Table 2.3, where the number of covariates fitted dropped precipitously as the frequency threshold got more stringent. If these rare haplotype alleles were not removed from the analysis, it is unlikely that prediction accuracy would rival that of the SNP model because most rare covariates will not explain much of the genetic variance due to their low frequency and will therefore be shrunk to zero (Gianola, 2013).

Prediction accuracies of haplotype models that used haploblocks of 500 kb or shorter (<8 SNPs per haploblock on average) generally had higher prediction accuracies than the SNP model, particularly when haplotype alleles with frequency less than 1% were removed from the training population. Models that fitted 125 kb haplotype alleles generally had higher prediction accuracy than the SNP model, indicating that there is a benefit of using these small haploblocks (2 SNPs per block on average) for genomic prediction rather than SNPs. Previous studies have evaluated the performance of haploblocks defined by the number of SNPs (e.g. 2 SNPs or 4 SNPs per haploblock), mostly using simulated data. Simulation studies using a similar density to our study found the optimal haploblock length ranged from 5-10 SNPs per haploblock (Villumsen and Janss, 2009; Villumsen et al., 2009) – slightly larger than the haploblock length that gave the highest prediction accuracy in our population.

Villumsen et al. (2009) demonstrated by simulation that the optimal number of SNPs in a haploblock depends on the distance between markers, the extent of LD and the population structure. It has also been shown by simulation that the purpose of the analysis can dictate the optimal haploblock length, with shorter haploblocks of 0.04-0.07 cM being preferable for QTL mapping and longer haploblocks of 0.12 cM, the longest haploblocks tested, performing better for genomic prediction (Calus et al., 2009). The optimal haplotype length to use for an analysis needs to be evaluated for each data set independently, with the purpose of the analysis (i.e. QTL mapping or genomic prediction) in mind.

Haplotype Allele Frequency Threshold

When using ~50k density SNPs to create haplotypes, the number of covariates to estimate is often much higher than the number of SNPs, which increases computation time, as seen in Table 2.3. The number of covariates that need to be estimated can be reduced by removing SNPs before generating haplotype alleles (Calus et al., 2009; Cuyabano et al., 2015b) or by only fitting covariates for haplotype alleles in regions that have known or putative QTL, along with a residual polygenic effect (Boichard et al., 2012; Cuyabano et al., 2015b). When appropriate filtering is performed, the resulting accuracy of genomic prediction can be equal to, or even higher than, genomic prediction using all haplotype alleles, as shown by Cuyabano et al. (2015b).

When haplotype alleles are fitted as random effects, as in BayesA, BayesB and BayesN, the estimated effects are shrunk relative to the variance assumed for that allele (i.e. $\sigma_e^2/\sigma_{\alpha_j}^2$) (Meuwissen et al., 2001; Gianola, 2013). A haplotype allele that is in low frequency will be shrunk more than another allele with a similar effect that is in moderate frequency. As expected, due to the polygenic nature of the traits in this study, removal of rare haplotypes for the smaller

haploblocks had little impact on prediction accuracy for frequency thresholds below 5% and haploblocks 500 kb or shorter, confirming that filtering based on haplotype allele frequency is a suitable method to decrease computation time (Table 2.3) with little loss in accuracy when haploblocks are an appropriate length for the data set.

Haplotype Models

The Hap250-1 models performed the best across the three traits investigated in this study but took much longer to run than the SNP models due to fitting almost double the number of covariates (Table 2.5). The Hap125-10 models fitted a similar number of covariates as the SNP models, and therefore had a similar runtime (Table 2.5), with a similar or higher accuracy (Figure 2.2). Although the Hap125-10 models tended to have lower accuracy than the Hap250-1 models, this difference was not significant for most breeds and traits (Figure 2.2). Therefore, fitting small haplotypes and removing rare haplotype alleles based on a high frequency threshold has the potential to improve prediction accuracy compared to fitting SNPs without the increased computation time and resources that are typically associated with haplotype analyses.

Bayesian Models

Genomic prediction accuracy has been shown to depend on the genetic architecture of the trait and whether prior assumptions of the model appropriately account for the number of loci that affect the trait and the distribution of their effects (Daetwyler et al., 2010; Hayes et al., 2010). BayesA (Meuwissen et al., 2001) was chosen to identify the impact of haploblock length on genomic prediction accuracy because it is provides a higher prediction accuracy than the Bayesian equivalent of GBLUP, BayesC0 (Kizilkaya et al., 2010), when a trait has large effect QTL (Meuwissen et al., 2001), such as have been identified for Fat and Lwt (Grisart et al., 2002; Karim et al., 2011). Although SCS is known as a very polygenic trait (Meredith et al., 2012),

suggesting that BayesC0 may be more appropriate, Habier et al. (2011) found that BayesA gave a higher predictive accuracy than GBLUP for SCS in North American Holstein bulls. Therefore BayesA was expected to be a suitable model for all traits evaluated in this study.

Cuyabano et al. (2015a) obtained higher prediction accuracy when fitting haplotype alleles rather than SNPs in genomic prediction models when using the Bayesian mixture model called BayesR (Erbe et al., 2014) – however this improvement was not observed when fitting a Bayesian GBLUP model. BayesR assumes that marker effects come from a mixture of four normal distributions, such that most markers have little or no effect (i.e. are sampled from a distribution with small variance), while a proportion of markers have a large effect (i.e. are sampled from a distribution with large variance). These results suggest that it is not appropriate to assume that haplotype allele effects follow a single normal distribution, such as in BayesC0, and further supports our use of a BayesA, where marker effects were assumed to have a marker-specific variance.

The BayesB and BayesN models were also evaluated in this study to determine which model would be more suitable for haplotype analyses and whether either model outperformed BayesA. When a large proportion of the variation in a trait is explained by few large QTL, BayesA, which estimates non-zero effects for all markers, has been shown to be less desirable than models such as BayesB, which estimate non-zero effects for a proportion of markers (Meuwissen et al., 2001). Of particular relevance to our study, a BayesB model with a high π value (i.e. estimating non-zero effects for very few markers each iteration) has been shown to outperform BayesA for traits such as milk fat yield when fitting SNPs (Habier et al., 2011). In our study, two Bayesian mixture models were evaluated in addition to BayesA: BayesB, which samples each haplotype allele independent of genomic region and BayesN, which samples

haplotype alleles within a genomic region jointly, based on whether or not the region is sampled that iteration. As implemented in our study, the BayesN haploblock model can be thought of as analogous to BayesB, where the haploblock is sampled as being associated with the trait or not, rather than the haplotype allele.

Performance of Different Bayesian Models

BayesB and BayesN were tested with a range of values for π and Π, respectively (Figures S2.6 and S2.7) and the following results were consistent between fitting covariates for haplotype alleles and covariates for SNPs. The prediction accuracy was very similar for pi values between 0 and 0.8, with slight variations such that the maximum accuracy was obtained at different pi values for each trait and breed. Almost all traits and breeds had a sharp decrease in accuracy when pi was greater than 0.8 for both BayesB and BayesN. This suggests that, fitting covariates for approximately 20% of the genome accounts for the effects of large QTL affecting the trait as well as the polygenic portion of the trait, likely through the genomic relationships between animals (Habier et al., 2007).

A minimal and non-significant difference in accuracy was observed between BayesA, BayesB and BayesN, provided appropriate pi values were used. Thus, all three methods are appropriate for genomic prediction fitting covariates for haplotype alleles (Figure 2.2). Consistent with results from Zeng et al. (2015) for this SNP density, fitting 2 SNPs per window in a BayesN SNP model resulted in slightly lower prediction accuracy than fitting all SNPs per window (Table S2.5). It was, however, surprising that BayesN did not have higher prediction accuracy than BayesB for haplotype models; conceptually, covariates with non-zero effects estimated in an iteration are more likely to be associated with the trait in BayesN because all haplotype alleles within a haploblock are included or excluded simultaneously. In contrast,

associations from BayesB analyses are more likely to be spurious because each haplotype allele independently has a zero or non-zero estimate sampled. BayesN also had higher bias than BayesB and BayesA models; however this was not significant in most cases and bias was comparable to the BayesA SNP model when Hap250-1 haplotypes were fit (Table 2.4).

Computing Time

BayesB had the shortest computing times of the Bayesian models that were tested in our study, followed by BayesA (Table 2.5). BayesN had much longer computing times that the other two methods which was not consistent with the finding from Zeng et al. (2015), where BayesN had a similar runtime to BayesB. When our data was tested using the C++ BayesN code used by Zeng et al. (2015), runtimes similar to BayesA, but longer than BayesB, were obtained. Thus, it may be possible to further improve runtime of BayesN when fitting covariates for haplotype alleles as implemented in our study by fixing $\delta_{ij} = 1$ and only sampling Δ, rather than sampling δ_{ij} for each haplotype allele (with probability $1 - \pi = 1$).

Models that fitted haplotype alleles typically fitted a larger number of covariates than models that fitted SNPs and therefore had longer runtimes. The development of a haplotype model for use in genomic prediction is appealing given the improvement in prediction accuracy when fitting haplotype alleles rather than SNPs. The BayesB Hap250-1 model had similar runtimes as the BayesA SNP model (Table 2.5) and equivalent or higher prediction accuracy for all traits (Figure 2.2).

Potential Impact of Haplotype Models on Selection Decisions

Theoretically, improvements in accuracy will result in improved genetic gain in a population (Falconer and Mackay, 1996); however, if this increased accuracy does not change

the ranking of individuals, it is unlikely to have a substantial impact on realized genetic gain. The Spearman rank correlation between the BayesA Hap250-1 and SNP models was evaluated to test whether fitting haplotype alleles rather than SNPs impacts selection decisions and genetic gain. The rank correlation when considering all animals was high (Table 2.6), suggesting no major re-ranking occurred and there were no cases where a cow had a particularly high DGV for the SNP model and a low DGV for the haplotype model, or vice versa.

In practice, only a small percentage of cows are selected to be dams of the next generation of sires (DairyNZ, 2009). Thus, re-ranking amongst the top cows may impact which individuals are selected as parents of the next generation. The rank correlation of the top 100 cows from either the SNP or Hap250-1 models was evaluated and found to be much lower than the rank correlation across all animals (Table 2.6). This corresponded with substantial differences in which of the cows would be selected as the top 0.9%. Considering the re-ranking of the top animals and the improvement in accuracy for haplotype models over SNP models that was observed in our study, genomic prediction fitting haplotype alleles is expected to result in higher realized genetic gain than genomic prediction fitting SNPs.

SNP Density

Increasing SNP density will influence the ability to differentiate sequence-resolution haplotype alleles within a haploblock: at sequence level all true haplotype alleles in the data set will theoretically be able to be identified; while at lower densities a single identified haplotype allele may represent two or more true haplotype alleles. This impacts the ability of a model to accurately estimate the BV of an animal for that haploblock because the effect of the identified haplotype alleles will be a weighted average of the effects of the underlying true haplotype

alleles, in addition to prediction error. Incorporating genotype at causal mutations into haplotypes will allow more accurate estimation of haplotype effects compared to not having the causal mutations in the haplotype, and improve the ability to detect short-range epistatic effects between loci contained within the same haploblock. Therefore, increasing SNP density has the potential to improve genomic prediction accuracy when using haplotype models. However, increasing SNP density will increase the number of identified haplotype alleles (Schopen and Schrooten, 2013) which will increase the number of rare haplotype alleles at a locus, shrinking the effect of these alleles toward zero (Gianola, 2013). This can potentially limit any improvement in prediction accuracy that would otherwise been seen from increasing SNP density.

Impact of Training Set

Training Set Size

Prediction accuracy has been shown to decline when the size of the training data set decreases (Karaman et al., 2016). Haplotype models likely are more sensitive to decreases in training data sizes because haplotype alleles that are present in a validation animal are less likely to be observed in a small training data set than a larger training data set. Haplotype allele effects can only be estimated for alleles that are observed in the training data set, so validation animals with many missing haplotype alleles are unlikely to be predicted with high accuracy. It is expected that at least 1,000 phenotypic records are needed to accurately estimate each haplotype allele effect (Hayes et al., 2007).

The number of animals in the training set may also impact the optimal haploblock length: a small training data set may result in shorter optimal haploblock lengths than a larger training

set. The ability of a haplotype model to accurately obtain DGV depends on both the power to accurately estimate the effect of the haplotype alleles fitted in the model, as well as the ability of those haplotype alleles to capture QTL effects and relationships between animals. Longer haploblocks generate a larger number of haplotype alleles than shorter haploblocks, and many of these are at low frequency in the population (Table 2.3) and therefore there is little power to detect associations when the training data set is small. Longer haplotypes also primarily capture more recent relationships – although if they get too long the relationship between parent and offspring or between full-sibs can be less than 0.5 (Ferdosi et al., 2016).

Multi-Breed Training Set

Our study used a training population consisting of multiple breeds, as is done in New Zealand genomic evaluations (Winkelman et al., 2015). Training on each breed separately may lead to higher accuracy in some cases, for example if the phase of a tagging SNP and large QTL differs by breed, or if some QTL only segregate in one breed (Saatchi et al., 2014). Fitting covariates for haplotypes rather than SNPs may improve genomic prediction accuracy by capturing breed-specific effects if haplotype alleles around these QTL are specific to a breed. Kachman et al. (2013) found that a training dataset that contained multiple beef breeds did not improve accuracy of genomic prediction using SNPs over a training dataset that contained the subset of animals that were of the same breed as the validation dataset. However, a combined training set of Danish, Swedish and Finnish Red cattle was found to increase genomic prediction accuracy using both SNPs (Brondum et al., 2011) and haplotypes (Cuyabano et al., 2015a), compared to within-breed training and validation datasets. These studies (Brondum et al., 2011; Kachman et al., 2013; Saatchi et al., 2014; Cuyabano et al., 2015a) suggest that the relationship between breeds, particularly around QTL, is an important factor in the success of genomic

prediction using a multi-breed training set. de Roos et al (2008) evaluated the genomic relationship between New Zealand Holstein Friesian, New Zealand Jersey and populations of Holsteins from the Netherlands and Australia. They found that phase was highly correlated among Holstein Friesians and Jerseys in New Zealand – higher than between New Zealand Holstein Friesians and their other Holstein populations. – indicating that it is appropriate to use a multi-breed training data set for genomic prediction of New Zealand dairy cattle.

Phasing Accuracy

Performance of haplotype models depends on the ability to accurately phase the genotypes of training and validation animals because phasing errors will result in the generation of incorrect haplotype alleles. Animals that are closely related are expected to share more haplotype alleles than animals that are distantly related (Ferdosi et al., 2016). Thus, phasing accuracy is expected to be higher in data sets that contain closely related animals than data sets with only distantly related animals (Weng et al., 2014). Phasing methods, such as LINKPHASE3 (Druet and Georges, 2015), that take advantage of pedigree information can improve phasing accurately, particularly when there are close relationships between animals in the dataset, i.e. sire and multiple offspring. The data set used for phasing in our study included over 58,000 animals, including most sires used in New Zealand in the past 20 years, as well as pedigree information confirmed through genotyping. These animals were initially phased using pedigree information, then any regions for which the phase was not clear were phased using population haplotypes from BEAGLE, as described in Druet and Georges (2015). Phasing accuracy is expected to be high in our data set because it is a large data set of animals that are closely related to others in the data set and the use of a method that takes advantage of pedigree information.

Fixed vs. Variable-length Haplotypes

Our study evaluated haplotypes that were based on a fixed-length, in megabases, throughout the genome. It has been shown that recombination rates vary across the genome in many species (Nachman, 2002) and that this variation is particularly large in dairy cattle (Sandor et al., 2012), suggesting that the optimal haploblock length for genomic prediction may differ across the genome because recombination breaks down LD and can create new haplotype alleles. Another reason why optimal haploblocks length may differ across the genome in domesticated plants and animals is because they have undergone artificial selection for production traits for many generations (Mignon-Grasteau et al., 2005), which has resulted in some regions around production-related QTL undergoing selective sweeps (Maynard-Smith and Haigh, 1974). This increases LD in these regions because haplotypes that capture the favorable QTL allele also increase in frequency (Wiener et al., 2003; Palaisa et al., 2004; Hayes et al., 2009). Thus, the optimal haploblock length may be longer around QTL that are being selected on and the optimal haploblock length when fitting haplotypes around QTL, as in Boichard et al. (2012), may be longer than was observed in our study. Methods to define haploblocks that take different recombination rates or LD patterns across the genome into account, termed "variable-length" haploblocks, may result in higher genomic prediction accuracy than fixed-length haploblocks.

Although variable-length haploblocks have the potential to improve genomic prediction accuracy over fixed-length haploblocks, they are often more complicated to generate. A number of methods to define the bounds of variable-length haploblocks from SNPs have been proposed, such as pairwise LD (Cuyabano et al., 2015a; Cuyabano et al., 2015b) or Identity-By-Descent (IBD) probabilities (Calus et al., 2008; Calus et al., 2009). These methods are more time

consuming than fixed-length methods because they additionally involve the calculation of LD or IBD probabilities.

Most commercial dairy cattle breeding programs use a small number of sires mated to a large number of females, so breeders need to be mindful of the relationship between commercial sires both within and across years to control levels of inbreeding in the population (Wray and Goddard, 1994). There is a low correlation between haplotype allele frequency in successive groups of sires compared to successive groups of dams, likely because so few sires are selected each generation and efforts are made to limit the relationships between common sires (de Roos et al., 2008). Haplotype alleles that are present in selected sires will be spread widely throughout the population in subsequent generations, which can rapidly increase the frequency of rare haplotypes alleles if they exist in a selected sire, influencing population-wide patterns of LD. If LD patterns change from generation-to-generation, variable-length haploblocks will likely have limited success when individuals that are being predicted are born multiple generations after individuals in the training population. Although variable-length haploblocks have the potential to capture more genetic variance, fixed-length haploblocks may do just as well if LD patterns quickly change over generations.

2.6 Conclusions

Fitting covariates for fixed-length haplotype alleles rather than SNPs can increase the accuracy of genomic prediction up to 5.5%. Haplotype length and filtering based on haplotype allele frequency have a large impact on prediction accuracy and bias and are therefore important parameters to optimize for the population and analysis that is being performed. In this data set, shorter haploblocks (125-250 kb; average of 2-4 SNPs per haploblock) resulted in higher

accuracies and generally lower biases than longer haploblocks (1 Mb or longer; average of at least 15 SNPs per haploblock), which had lower accuracies than the SNP model and deemed too long for genomic prediction in the New Zealand dairy cattle population. A more stringent haplotype allele frequency filter tended to decrease prediction accuracy, particularly when haploblocks were long. The BayesA model that consistently gave the highest accuracy and lowest bias was the model that fitted 250 kb haploblocks with a 1% haplotype allele frequency filter.

The Bayesian model that was used for haplotype models (BayesA, BayesB or BayesN) had very little impact on prediction accuracy, as long as π and Π values were less than 0.8 for the BayesB and BayesN models. Fitting 125 kb haplotypes with a 10% frequency filter resulted in equivalent or higher prediction accuracy than fitting SNPs. The BayesA model that fitted 250 kb haplotypes with a 1% frequency filter performed well for all traits, improving accuracy up to 5.5% compared to the BayesA SNP model across breeds and traits. The BayesB model fitting 250 kb haplotype alleles present in the training data set at >1% frequency had similar accuracy and bias as BayesA and BayesN models but a much shorter runtime. Comparing the ranking of the top animals in SNP model to the haplotype model suggested that the improvement in accuracy from utilizing haplotype models would result in a difference in which individuals are selected as parents of the next generation. Further studies should assess the impact of constructing haplotypes that better capture the population structure; as such methods may result in a greater improvement in genomic prediction models.

2.7 Acknowledgements

The authors would like to thank Kathryn Tiplady and Dr. Bevin Harris from Livestock Improvement Corporation for providing the Yield Deviation phenotypes. The authors would also like to thank Dr. Marcos Barbosa da Silva, Dr. Jack Dekkers, Dr. Xiaochen Sun and Dr. Jian Zeng for their discussions on haplotype analyses and Bayesian modeling. Tom Druet is Research Associate from the Fonds de la Recherché Scientifique- FNRS (F.R.S.-FNRS).

2.8 Author Contributions

MH designed and ran the analyses, interpreted the results and wrote the manuscript. TD phased the genotypes and critically contributed to the manuscript. AH assisted with the study design, interpretation of results and critically contributed to the manuscript. DG supervised the study and critically contributed to the manuscript.

2.9 References

Arias, J. A., M. Keehan, P. Fisher, W. Coppieters, and R. Spelman. 2009. A high density linkage map of the bovine genome. Bmc Genetics 10doi: 10.1186/1471-2156-10-18

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Animal Production Science 52(2-3):115-120. (Review) doi: 10.1071/an11119

Brondum, R. F., E. Rius-Vilarrasa, I. Stranden, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. Journal of Dairy Science 94(9):4700-4707. (Article) doi: 10.3168/jds.2010-3765

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81(5):1084-1097. doi: 10.1086/521987

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178(1):553-561. (Article) doi: 10.1534/genetics.107.080838

Calus, M. P. L., T. H. E. Meuwissen, J. J. Windig, E. F. Knol, C. Schrooten, A. L. J. Vereijken, and R. F. Veerkamp. 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. Genetics, Selection, Evolution 41(11):(15 January 2009). (article)

Cuyabano, B. C. D., G. Su, G. J. M. Rosa, M. S. Lund, and D. Gianola. 2015a. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. Journal of Dairy Science 98(10):7351-7363. doi: 10.3168/jds.2015-9360

Cuyabano, B. C. D., G. S. Su, and M. S. Lund. 2015b. Selection of haplotype variables from a high-density marker map for genomic prediction. Genetics Selection Evolution 47:11. (Article) doi: 10.1186/s12711-015-0143-3

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185(3):1021-1031. doi: 10.1534/genetics.110.116855

DairyNZ. 2009. New Zealand Dairy Herd Improvement Database Review. In: DairyNZ (ed.) http://www.dairynz.co.nz/media/532738/anderson_report.pdf.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179(3):1503-1512. (Article) doi: 10.1534/genetics.107.084301

de Roos, A. P. W., C. Schrooten, and T. Druet. 2011. Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. Journal of Dairy Science 94(9):4708-4714. doi: 10.3168/jds.2010-3905

Druet, T., and M. Georges. 2010. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184(3):779-798. (Article)

Druet, T., and M. Georges. 2015. LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. Bioinformatics 31(10):1677-1679. doi: 10.1093/bioinformatics/btu859

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2014. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 97(10):6622-6622. doi: 10.3168/jds.2014-97-10-6622

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. Introduction to quantitative genetics. (Ed. 4):xv + 464 pp. (Book)

Ferdosi, M. H., J. Henshall, and B. Tier. 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genetics Selection Evolution

Garrick, D., and R. Fernando. 2013. Implementing a QTL Detection Study (GWAS) Using Genomic Prediction Methodology, Genome-Wide Association Studies and Genomic Prediction. Springer. p. 275-298.

Gianola, D. 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194(3):573-596. (Article) doi: 10.1534/genetics.113.151753

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12(2):222-231. (Article) doi: 10.1101/gr.224202

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389-2397. (Article) doi: 10.1534/genetics.107.081190

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. Bmc Bioinformatics 12:12. (Article) doi: 10.1186/1471-2105-12-186

Harris, B. L. 2005. Breeding dairy cows for the future in New Zealand. New Zealand Veterinary Journal 53(6):384-389. (Review) doi: 10.1080/00480169.2005.36582

Hayes, B. J., A. J. Chamberlain, S. Maceachern, K. Savin, H. McPartlan, I. MacLeod, L. Sethuraman, and M. E. Goddard. 2009. A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. Animal Genetics 40(2):176-184. doi: 10.1111/j.1365-2052.2008.01815.x

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetics Research 89(4):215-220. (Article) doi: 10.1017/s0016672307008865

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. Plos Genetics 6(9)doi: 10.1371/journal.pgen.1001139

Heidaritabar, M., M. P. L. Calus, H. J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics

Hickey, J. M., B. P. Kinghorn, B. Tier, S. A. Clark, J. H. J. van der Werf, and G. Gorjanc. 2013. Genomic evaluations using similarity between haplotypes. Journal of Animal Breeding and Genetics 130(4):259-269. doi: 10.1111/jbg.12020

Kachman, S. D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn, W. M. Snelling, R. M. Thallman, M. Saatchi, D. J. Garrick, R. D. Schnabel, J. F. Taylor, and E. J. Pollak. 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genetics Selection Evolution 45:9. (Article) doi: 10.1186/1297-9686-45-30

Karaman, E., H. Cheng, M. Firat, D. Garrick, and R. Fernando. 2016. An upper bound for accuracy of prediction using GBLUP. PLOS One

Karim, L., H. Takeda, L. Lin, T. Druet, J. A. C. Arias, D. Baurain, N. Cambisano, S. R. Davis, F. Farnir, B. Grisart, B. L. Harris, M. D. Keehan, M. D. Littlejohn, R. J. Spelman, M. Georges, and W. Coppieters. 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nature Genetics 43(5):405-+. (Article) doi: 10.1038/ng.814

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. Journal of Animal Science 88(2):544-551. (Article) doi: 10.2527/jas.2009-2064

LIC. 2009. Your Index Your Animal Evaluation System.

LIC, and DairyNZ. 2015. New Zealand Dairy Statistics 2014-15, http://www.dairynz.co.nz/media/3136117/new-zealand-dairy-statistics-2014-15.pdf.

Matukumalli, L., S. Schroeder, S. DeNise, T. Sonstegard, C. Lawley, M. Georges, W. Coppieters, K. Gietzen, J. Medrano, and G. Rincon. 2011. Analyzing LD blocks and CNV segments in cattle: novel genomic features identified using the BovineHD BeadChip. San Diego, CA: Illumina Inc

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. Plos One 4(4):13. (Article) doi: 10.1371/journal.pone.0005350

Maynard-Smith, J., and J. Haigh. 1974. Hitch-hiking effect of a favorable gene. Genetics Research 23(1):23-35. (Article) doi: 10.1017/s0016672300014634

Meredith, B. K., F. J. Kearney, E. K. Finlay, D. G. Bradley, A. G. Fahey, D. P. Berry, and D. J. Lynn. 2012. Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. Bmc Genetics 13doi: 10.1186/1471-2156-13-21

Meuwissen, T., and M. Goddard. 2010. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics 185(2):623-U338. (Article) doi: 10.1534/genetics.110.116590

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating Improvement of Livestock with Genomic Selection. In: H. A. Lewin and R. M. Roberts, editors, Annual Review of Animal Biosciences, Vol 1. Annual Review of Animal Biosciences No. 1. Annual Reviews, Palo Alto. p. 221-237.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829. (Article)

Mignon-Grasteau, S., A. Boissy, J. Bouix, J. M. Faure, A. D. Fisher, G. N. Hinch, P. Jensen, P. Le Neindre, P. Mormede, P. Prunet, M. Vandeputte, and C. Beaumont. 2005. Genetics of adaptation and domestication in livestock. Livestock Production Science 93(1):3-14. doi: 10.1016/j.livprodsci.2004.11.001

Nachman, M. W. 2002. Variation in recombination rate across the genome: evidence and implications. Current Opinion in Genetics & Development 12(6):657-663. doi: 10.1016/s0959-437x(02)00358-1

Palaisa, K., M. Morgante, S. Tingey, and A. Rafalski. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. Proceedings of the National Academy of Sciences of the United States of America 101(26):9885-9890. doi: 10.1073/pnas.0307839101

Saatchi, M., R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. Bmc Genomics 15:16. (Article) doi: 10.1186/1471-2164-15-442

Sandor, C., W. B. Li, W. Coppieters, T. Druet, C. Charlier, and M. Georges. 2012. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. Plos Genetics 8(7):13. (Article) doi: 10.1371/journal.pgen.1002854

Schopen, G. C. B., and C. Schrooten. 2013. Reliability of genomic evaluations in Holstein-Friesians using haplotypes based on the BovineHD Bead Chip. Journal of Dairy Science 96(12):7945-7951. (Article) doi: 10.3168/jds.2012-6510

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. R. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (similar to 54,000) and high-density (similar to 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95(8):4657-4665. (Article) doi: 10.3168/jds.2012-5379

Sun, X., H. Su, P. Boddhireddy, and D. Garrick. 2016. Haplotype-based Genomic Prediction of Breeds Not in Training. In: Plant and Animal Genomes Conference XXIV, San Diego, CA

Turner, S., L. L. Armstrong, Y. Bradford, C. S. Carlson, Crawford, D.C., A. T. Crenshaw, M. Andrade, K. F. Doheny, J. L. Haines, G. Hayes, and G. Jarvik. 2011. Quality control procedures for genome-wide association studies. Current protocols in human genetics:1-19.

van Binsbergen, R., M. P. L. Calus, M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47:13. (Article) doi: 10.1186/s12711-015-0149-x

Van Raden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91(11):4414-4423. (Article) doi: 10.3168/jds.2007-0980

Van Raden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92(1):16-24. doi: http://dx.doi.org/10.3168/jds.2008-1514

Van Raden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal-model information. Journal of Dairy Science 74(8):2737-2746. doi: 10.3168/jds.S0022-0302(91)78453-1

Villumsen, T. M., and L. Janss. 2009. Bayesian genomic selection: the effect of haplotype length and priors. BMC proceedings 3 Suppl 1:S11.

Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics 126(1):3-13. (Article) doi: 10.1111/j.1439-0388.2008.00747.x

Weng, Z. Q., M. Saatchi, R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genetics Selection Evolution 46doi: 10.1186/1297-9686-46-34

Wiener, P., D. Burton, P. Ajmone-Marsan, S. Dunner, G. Mommens, I. J. Nijman, C. Rodellar, A. Valentini, and J. L. Williams. 2003. Signatures of selection? Patterns of microsatellite diversity on a chromosome containing a selected locus. Heredity 90(5):350-358. (Article) doi: 10.1038/sj.hdy.6800257

Winkelman, A. M., D. L. Johnson, and B. L. Harris. 2015. Application of genomic evaluation to dairy cattle in New Zealand. Journal of Dairy Science 98(1):659-675. doi: 10.3168/jds.2014-8560

Wray, N. R., and M. E. Goddard. 1994. Increasing long-term response to selection. Genetics Selection Evolution 26(5):431-451. doi: 10.1051/gse:19940504

Zeng, J. 2015. Whole genome analyses accounting for structures in genotype data, Iowa State University, http://lib.dr.iastate.edu/etd/14699.

Zondervan, K. T., and L. R. Cardon. 2004. The complex interplay among factors that influence allelic association. Nature Reviews Genetics 5(2):89-U14. (Review) doi: 10.1038/nrg1270

2.10 Figures

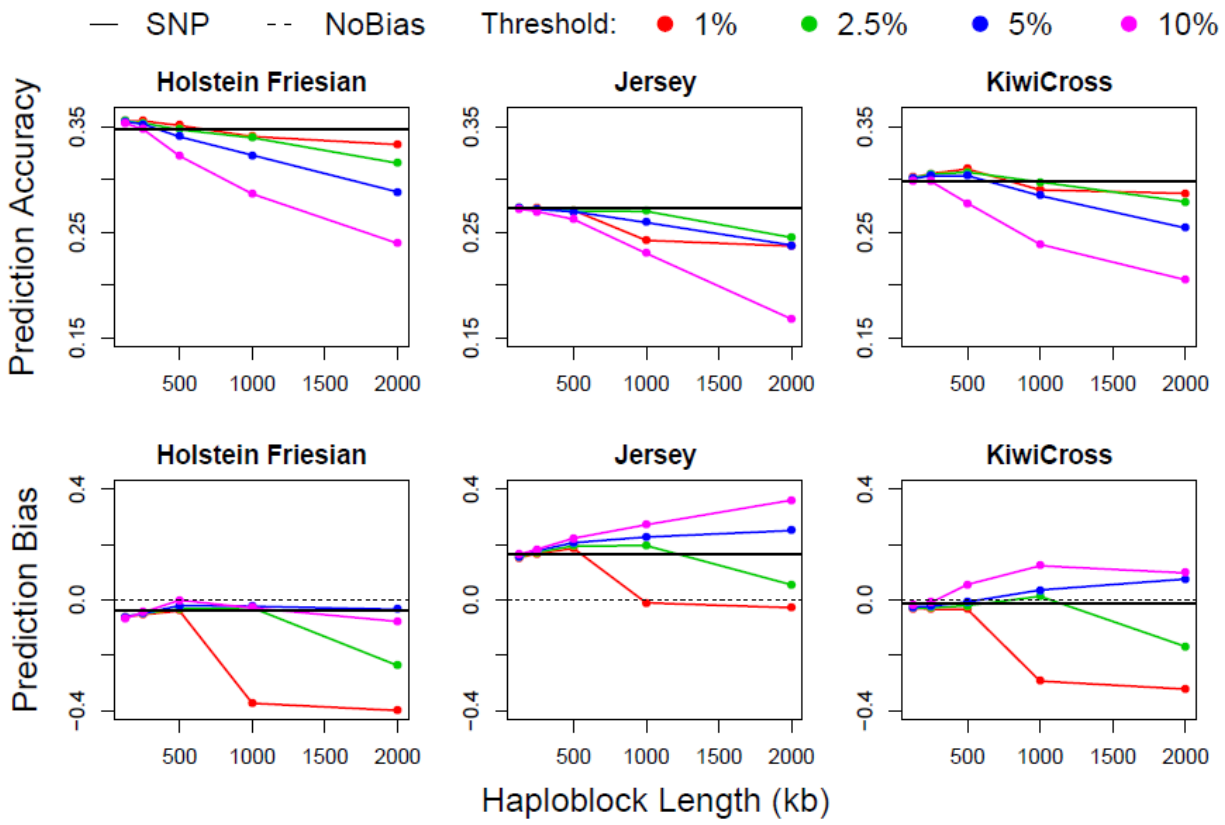


Figure 2.1: Genomic Prediction Accuracy and Bias of Milk Fat Yield with Varying Haplotype Lengths and Frequencies

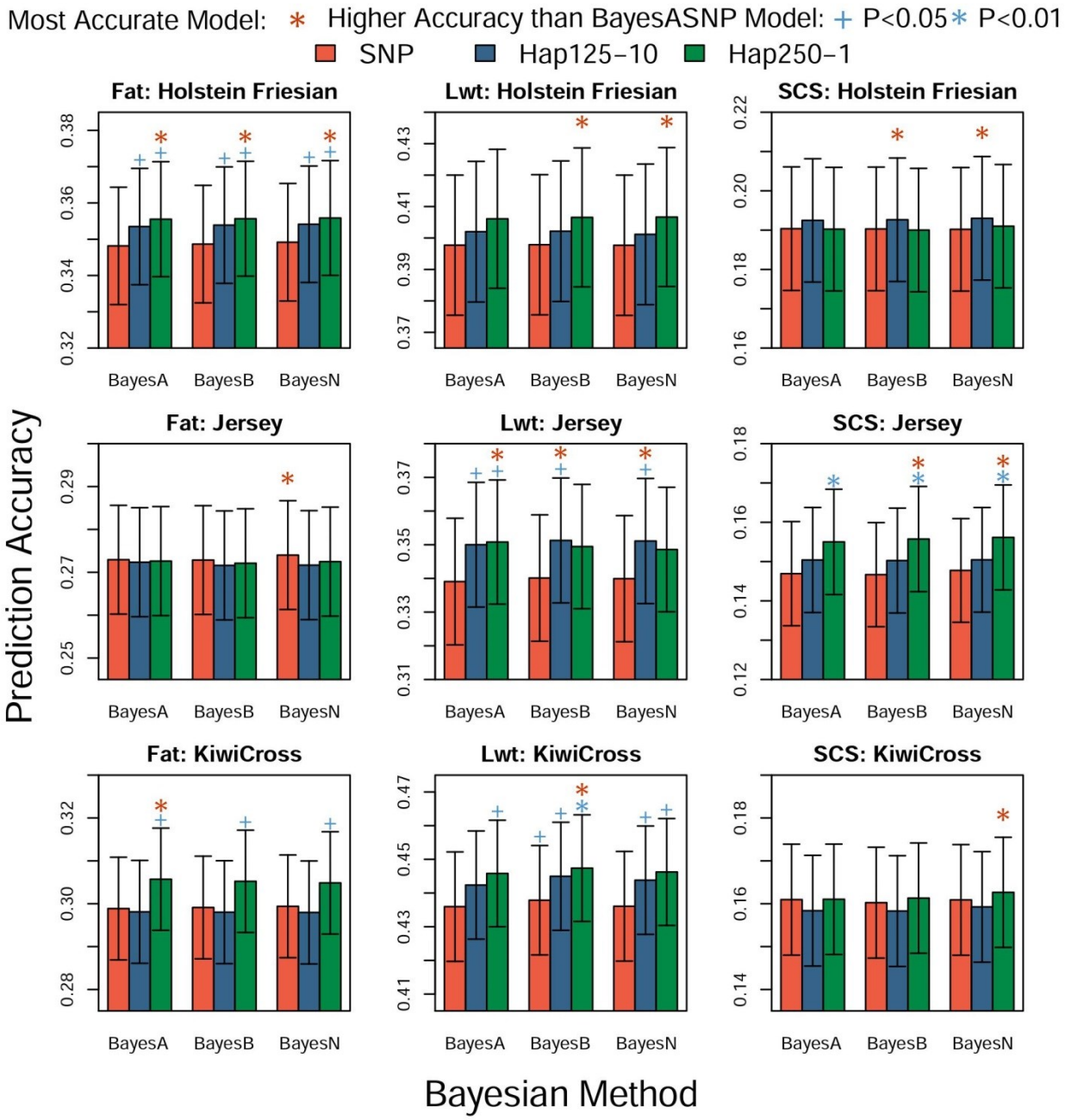
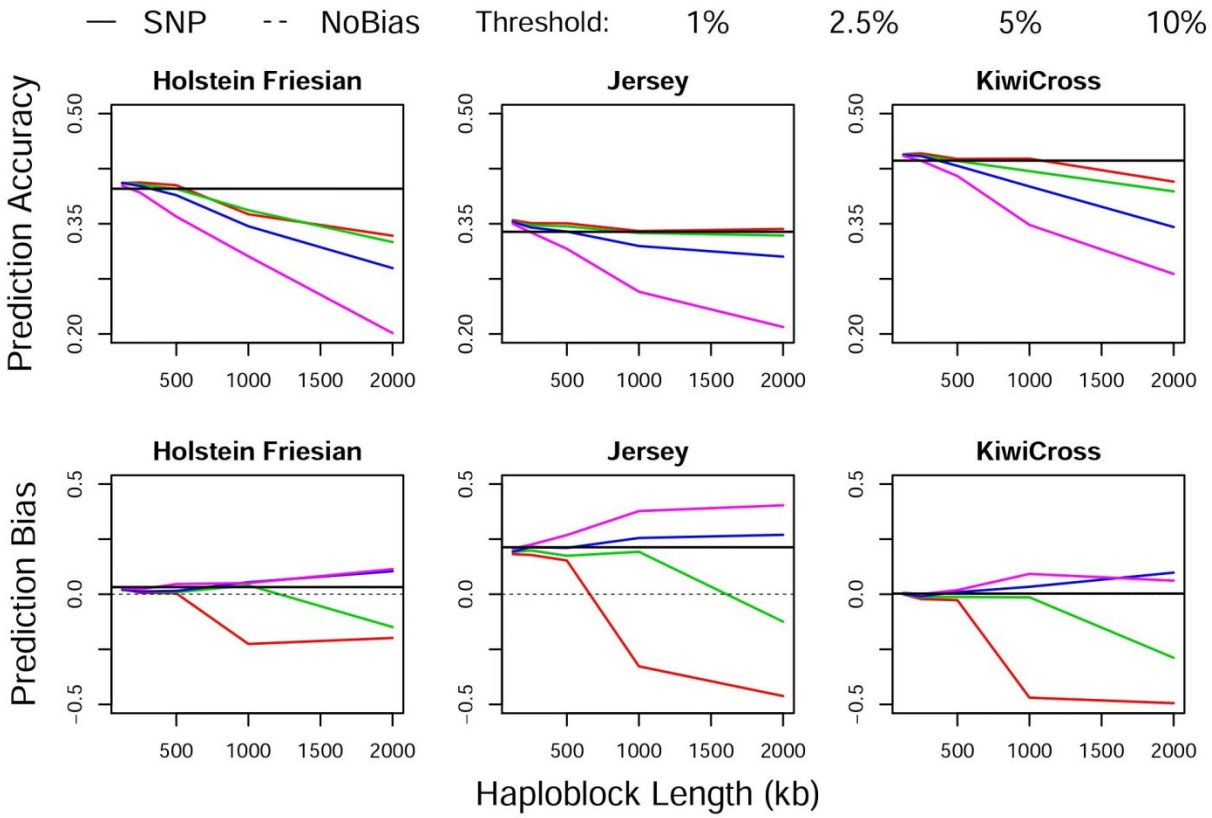
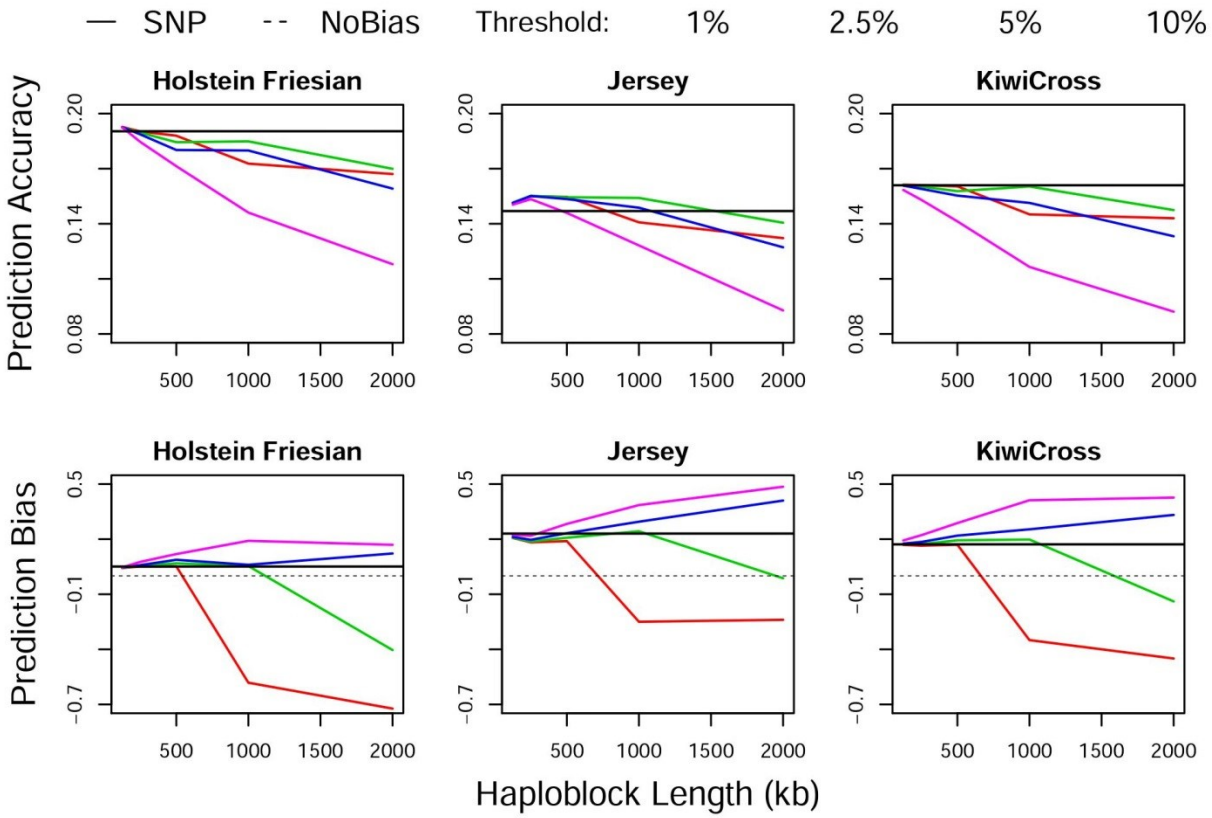


Figure 2.2: Genomic Prediction Accuracy of Bayesian SNP and Haplotype Models



Additional File S2.1: Genomic Prediction Accuracy and Bias of Liveweight with Varying Haplotype Lengths and Frequencies



Additional File S2.2: Genomic Prediction Accuracy and Bias of Somatic Cell Score with Varying Haplotype Lengths and Frequencies

Breed	Filter	Accuracy					Bias				
		125 kb	250 kb	500 kb	1 Mb	2 Mb	125 kb	250 kb	500 kb	1 Mb	2 Mb
Milk Fat Yield											
Holstein Friesian	1%	*0.356* *(0.016)*	*0.356* *(0.016)*	0.351 (0.016)	0.341 (0.016)	0.333 (0.016)	-0.063 (0.050)	-0.053 (0.049)	-0.040 (0.049)	*-0.373* *(0.067)*	*-0.399* *(0.071)*
	2.5%	*0.355* *(0.016)*	0.354 (0.016)	0.347 (0.016)	0.340 (0.016)	0.316 (0.017)	-0.062 (0.050)	-0.049 (0.049)	-0.032 (0.049)	-0.029 (0.051)	*-0.237* *(0.068)*
	5%	*0.355* *(0.016)*	0.353 (0.016)	0.341 (0.016)	0.323 (0.016)	0.288 (0.017)	-0.061 (0.050)	-0.049 (0.049)	-0.021 (0.050)	-0.023 (0.054)	-0.035 (0.063)
	10%	*0.353* *(0.016)*	0.348 (0.016)	0.322 (0.016)	0.286 (0.017)	0.240 (0.017)	-0.067 (0.051)	-0.045 (0.050)	-0.001 (0.052)	-0.029 (0.063)	-0.078 (0.078)
Jersey	1%	0.273 (0.013)	0.273 (0.013)	0.271 (0.013)	0.242 (0.013)	0.237 (0.013)	*0.152* *(0.040)*	*0.167* *(0.040)*	*0.185* *(0.039)*	-0.011 (0.054)	-0.029 (0.056)
	2.5%	0.273 (0.013)	0.272 (0.013)	0.270 (0.013)	0.270 (0.013)	0.245 (0.013)	*0.154* *(0.040)*	*0.172* *(0.040)*	*0.194* *(0.039)*	*0.196* *(0.039)*	0.054 (0.049)
	5%	0.273 (0.013)	0.272 (0.013)	0.270 (0.013)	0.259 (0.013)	0.238 (0.013)	*0.157* *(0.040)*	*0.178* *(0.039)*	*0.206* *(0.038)*	*0.227* *(0.039)*	*0.250* *(0.041)*
	10%	0.272 (0.013)	0.270 (0.013)	0.262 (0.013)	0.230 (0.013)	0.168 (0.013)	*0.164* *(0.040)*	*0.183* *(0.040)*	*0.222* *(0.039)*	*0.271* *(0.041)*	*0.360* *(0.050)*
KiwiCross	1%	0.302 (0.012)	*0.306* *(0.012)*	*0.310* *(0.012)*	0.290 (0.012)	0.287 (0.012)	-0.03 (0.042)	-0.033 (0.041)	-0.034 (0.041)	*-0.293* *(0.056)*	*-0.322* *(0.058)*
	2.5%	0.302 (0.012)	*0.305* *(0.012)*	*0.307* *(0.012)*	0.298 (0.012)	0.279 (0.012)	-0.029 (0.042)	-0.029 (0.041)	-0.021 (0.041)	0.013 (0.041)	*-0.169* *(0.053)*
	5%	0.301 (0.012)	0.304 (0.012)	0.304 (0.012)	0.285 (0.012)	0.254 (0.012)	-0.025 (0.042)	-0.023 (0.041)	-0.008 (0.041)	0.035 (0.042)	0.075 (0.046)
	10%	0.298 (0.012)	0.298 (0.012)	0.278 (0.012)	0.239 (0.012)	0.205 (0.012)	-0.017 (0.042)	-0.010 (0.042)	0.055 (0.042)	*0.123* *(0.047)*	0.097 (0.056)

Additional File S2.3: Prediction Accuracy and Bias for BayesA Haplotype Models

Liveweight											
Holstein Friesian	1%	0.406 (0.022)	0.406 (0.022)	0.402 (0.022)	0.363 (0.023)	0.334 (0.024)	0.026 (0.058)	0.006 (0.058)	0.004 (0.060)	*-0.226 (0.083)*	*-0.198 (0.089)*
	2.5%	0.406 (0.022)	0.403 (0.022)	0.397 (0.023)	0.368 (0.022)	0.325 (0.024)	0.024 (0.058)	0.011 (0.058)	0.009 (0.061)	0.041 (0.062)	-0.149 (0.085)
	5%	0.406 (0.022)	0.401 (0.022)	0.389 (0.023)	0.347 (0.023)	0.290 (0.024)	0.020 (0.058)	0.012 (0.058)	0.015 (0.062)	0.055 (0.065)	0.104 (0.075)
	10%	0.402 (0.022)	0.392 (0.022)	0.360 (0.023)	0.306 (0.023)	0.201 (0.026)	0.027 (0.058)	0.022 (0.059)	0.046 (0.065)	0.051 (0.076)	0.114 (0.111)
Jersey	1%	*0.355 (0.018)*	*0.351 (0.018)*	0.351 (0.019)	0.340 (0.019)	0.343 (0.019)	*0.182 (0.045)*	*0.178 (0.045)*	*0.153 (0.047)*	*-0.327 (0.078)*	*-0.462 (0.085)*
	2.5%	*0.353 (0.019)*	0.348 (0.018)	0.346 (0.018)	0.337 (0.018)	0.334 (0.019)	*0.191 (0.045)*	*0.199 (0.044)*	*0.174 (0.046)*	*0.193 (0.046)*	-0.125 (0.066)
	5%	*0.353 (0.019)*	0.345 (0.018)	0.339 (0.018)	0.320 (0.018)	0.305 (0.019)	*0.194 (0.045)*	*0.214 (0.044)*	*0.210 (0.045)*	*0.255 (0.045)*	*0.270 (0.046)*
	10%	*0.350 (0.019)*	0.338 (0.019)	0.316 (0.019)	0.257 (0.019)	0.209 (0.019)	*0.206 (0.044)*	*0.225 (0.044)*	*0.268 (0.045)*	*0.377 (0.047)*	*0.404 (0.056)*
KiwiCross	1%	*0.445 (0.016)*	*0.446 (0.016)*	0.438 (0.016)	0.439 (0.016)	0.407 (0.017)	-0.001 (0.041)	-0.022 (0.041)	-0.027 (0.042)	*-0.469 (0.060)*	*-0.494 (0.067)*
	2.5%	*0.444 (0.016)*	0.443 (0.016)	0.435 (0.016)	0.422 (0.016)	0.394 (0.017)	0.000 (0.041)	-0.014 (0.042)	-0.012 (0.042)	-0.014 (0.043)	*-0.288 (0.061)*
	5%	*0.444* (0.016)	0.442 (0.016)	0.429 (0.016)	0.401 (0.016)	0.346 (0.017)	0.002 (0.041)	-0.009 (0.041)	0.009 (0.042)	0.034 (0.044)	*0.098 (0.047)*
	10%	0.442 (0.016)	0.436 (0.016)	0.415 (0.016)	0.348 (0.017)	0.282 (0.018)	0.007 (0.041)	0.000 (0.042)	0.019 (0.043)	0.092 (0.048)	0.062 (0.065)

Additional File S2.3 (cont.): Prediction Accuracy and Bias for BayesA Haplotype Models

Somatic Cell Count											
Holstein Friesian	1%	0.193 (0.016)	0.190 (0.016)	0.188 (0.016)	0.173 (0.016)	0.167 (0.016)	0.043 (0.080)	0.051 (0.081)	0.052 (0.083)	*-0.582* *(0.154)*	*-0.723* *(0.175)*
	2.5%	0.192 (0.016)	0.190 (0.016)	0.184 (0.016)	0.185 (0.016)	0.170 (0.016)	0.045 (0.080)	0.054 (0.081)	0.069 (0.083)	0.054 (0.085)	*-0.404* *(0.139)*
	5%	0.193 (0.016)	0.189 (0.016)	0.180 (0.016)	0.180 (0.016)	0.159 (0.016)	0.045 (0.080)	0.058 (0.081)	0.088 (0.084)	0.060 (0.086)	0.122 (0.091)
	10%	0.192 (0.016)	0.185 (0.016)	0.171 (0.016)	0.146 (0.017)	0.118 (0.017)	0.046 (0.080)	0.076 (0.081)	0.119 (0.085)	0.191 (0.094)	0.169 (0.120)
Jersey	1%	*0.151* *(0.013)*	*0.155* *(0.013)*	*0.155* *(0.013)*	0.141 (0.014)	0.132 (0.014)	*0.206* *(0.069)*	*0.183* *(0.070)*	*0.189* *(0.069)*	*-0.250* *(0.118)*	-0.239 (0.127)
	2.5%	*0.152* *(0.013)*	*0.155* *(0.013)*	*0.154* *(0.013)*	0.154 (0.013)	0.141 (0.014)	*0.207* *(0.069)*	*0.187* *(0.069)*	*0.208* *(0.067)*	*0.243* *(0.064)*	-0.013 (0.097)
	5%	*0.151* *(0.013)*	*0.155* *(0.013)*	0.153 (0.013)	0.149 (0.013)	0.127 (0.013)	*0.212* *(0.069)*	*0.196* *(0.068)*	*0.232* *(0.065)*	*0.295* *(0.062)*	*0.410* *(0.061)*
	10%	0.150 (0.013)	*0.153* *(0.013)*	0.146 (0.013)	0.128 (0.013)	0.093 (0.013)	*0.224* *(0.068)*	*0.218* *(0.067)*	*0.283* *(0.064)*	*0.386* *(0.062)*	*0.485* *(0.072)*
KiwiCross	1%	0.161 (0.013)	0.161 (0.013)	0.160 (0.013)	0.145 (0.013)	0.143 (0.013)	*0.169* *(0.067)*	*0.165* *(0.067)*	*0.169* *(0.067)*	*-0.350* *(0.121)*	*-0.450* *(0.134)*
	2.5%	0.161 (0.013)	0.160 (0.013)	0.158 (0.013)	0.160 (0.013)	0.147 (0.013)	*0.171* *(0.067)*	*0.173* *(0.066)*	*0.194* *(0.066)*	*0.198* *(0.064)*	-0.139 (0.101)
	5%	0.161 (0.013)	0.159 (0.013)	0.155 (0.013)	0.151 (0.013)	0.133 (0.013)	*0.176* *(0.066)*	*0.185* *(0.066)*	*0.219* *(0.065)*	*0.254* *(0.063)*	*0.332* *(0.065)*
	10%	0.158 (0.013)	0.153 (0.013)	0.141 (0.013)	0.117 (0.013)	0.092 (0.013)	*0.193* *(0.066)*	*0.221* *(0.065)*	*0.288* *(0.065)*	*0.412* *(0.064)*	*0.426* *(0.080)*

Accuracy: Significantly better than SNP model (one-sided paired t-test)
Bias: Significantly different from zero (two-sided t-test)

Additional File S2.3 (cont.): Prediction Accuracy and Bias for BayesA Haplotype Models

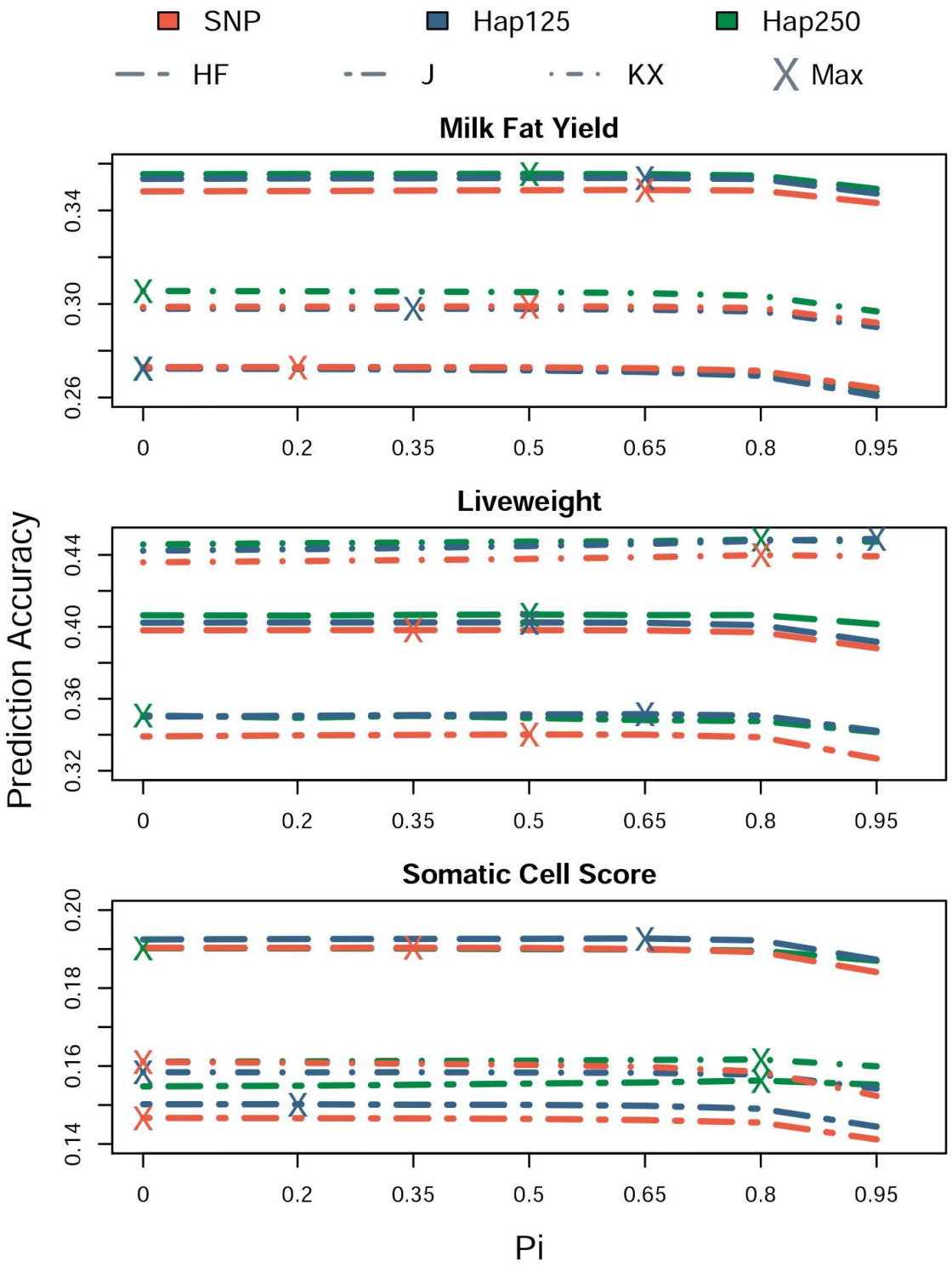
Trait	Filter	Mean Square Error				
		125 kb	250 kb	500 kb	1 Mb	2 Mb
Milk Fat Yield	SNP	611	611	611	611	611
	1%	610	609	610	614	615
	2.50%	611	610	611	614	617
	5%	611	610	612	619	625
	10%	611	612	619	630	640
Liveweight	SNP	888	888	888	888	888
	1%	882	881	883	909	919
	2.50%	882	884	886	901	917
	5%	881	885	893	916	945
	10%	884	892	911	956	993
Somatic Cell Score	SNP	1.26	1.26	1.26	1.26	1.26
	1%	1.26	1.26	1.26	1.27	1.27
	2.50%	1.26	1.26	1.26	1.26	1.27
	5%	1.26	1.26	1.26	1.27	1.28
	10%	1.26	1.26	1.27	1.28	1.29

Additional File S2.4: Mean Square Errors for the BayesA Models

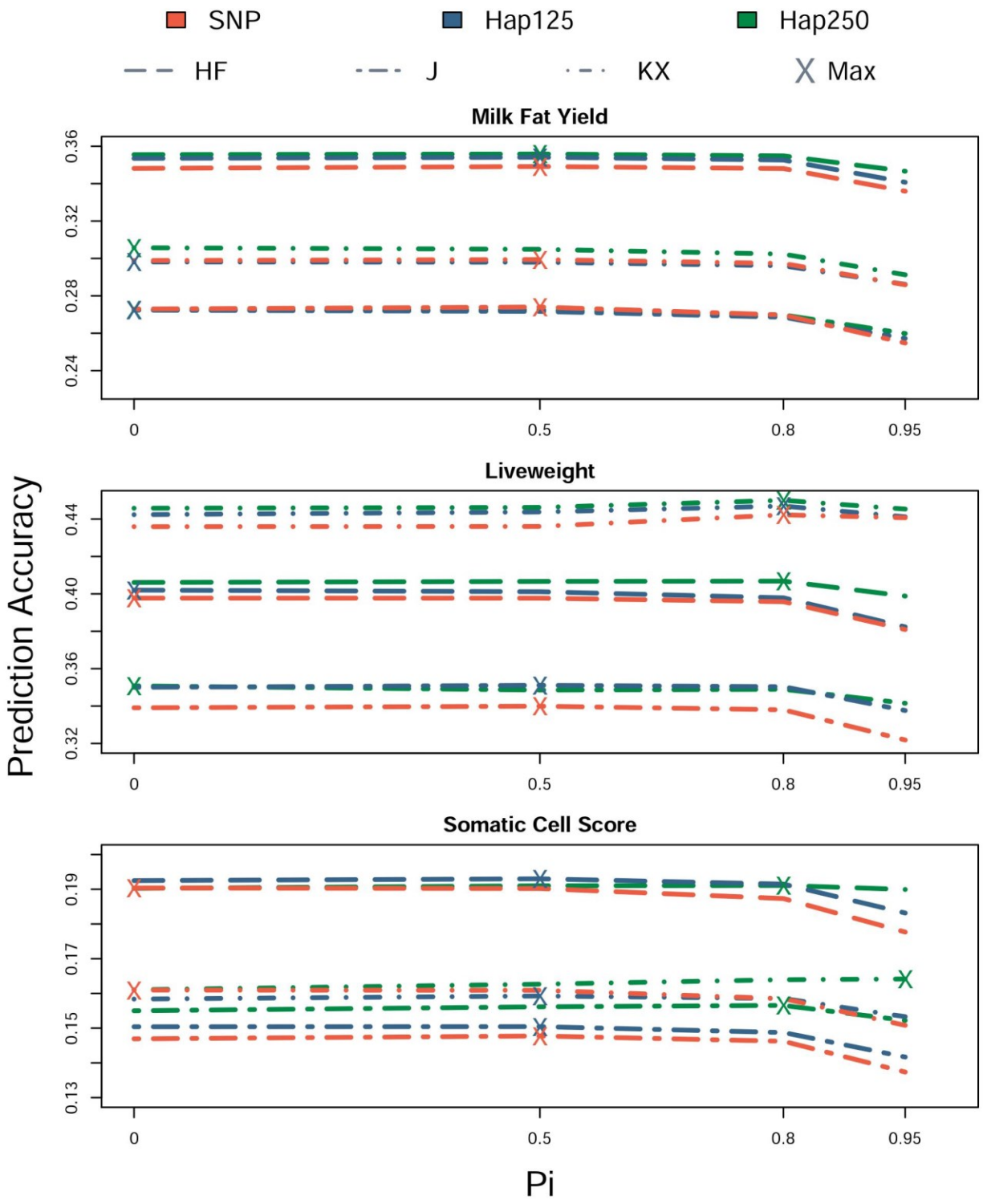
Trait[1]	Window Length	Holstein Friesian		Jersey		KiwiCross	
		2 SNPs	All SNPs	2 SNPs	All SNPs	2 SNPs	All SNPs
Fat	125 kb	0.349	0.349	0.27	0.273	0.301	0.3
	250 kb	0.348	0.349	0.271	0.274	0.3	0.299
	1 Mb	0.344	0.349	0.266	0.272	0.295	0.299
Lwt	125 kb	0.395	0.398	0.342	0.341	0.443	0.438
	250 kb	0.395	0.398	0.339	0.34	0.444	0.436
	1 Mb	0.388	0.396	0.331	0.345	0.442	0.444
SCS	125 kb	0.189	0.191	0.146	0.147	0.156	0.16
	250 kb	0.187	0.19	0.144	0.148	0.155	0.161
	1 Mb	0.184	0.19	0.142	0.147	0.151	0.16

1) Fat = Milk Fat Yield; Lwt = Liveweight; SCS = Somatic Cell Score

Additional File S2.5: Accuracy of BayesN Model with Π=0.5 Fitting 2 or All SNPs per Window



Additional File S2.6: Accuracy of BayesB Models with Varying π Values



Additional File S2.7: Accuracy of BayesN Models with Varying Π Values

CHAPTER III

COMPARISON OF HAPLOTYPE BLOCKING METHODS FOR GENOMIC PREDICTION IN AN ADMIXED POPULATION

Melanie Hayr[1,2], Andrew Hess[1], Jack Dekkers[1], and Dorian Garrick[1,3]

[1]Iowa State University, Iowa, USA

[2]LIC, Hamilton, New Zealand

[3]Massey University, Palmerston North, New Zealand

A paper to be submitted to *Genetics Selection Evolution*

3.1 Abstract

Background

Genomic prediction fitting haplotype alleles has been shown to improve prediction accuracy compared to SNPs; however the increase in accuracy depends on how haplotype blocks are defined. Genotypes on approximately 58,000 New Zealand dairy cattle at 37,740 SNPs from Illumina BovineSNP50 or HD panels were phased using LinkPHASE3 and DAGPHASE. Genotyped females born before 1 June 2008 were used for training, and genomic predictions for milk fat yield (n = 24,823), liveweight (n = 13,283) or somatic cell score (n = 24,864) were validated within breed (predominantly Holstein Friesian, predominantly Jersey, or admixed KiwiCross) in later-born genotyped females. The SNPs were assigned to approximately 9,670

haplotype blocks based on five methods: 1) Fixed length of 250 kb; 2) Pairwise linkage disequilibrium (LD) based on D'; 3) Multi-locus LD based on a G-test; 4) Identified recombination events, and 5) Number of haplotype alleles generated. Prediction accuracy and bias were evaluated using a BayesA model, fitting all SNPs or haplotype alleles, or haplotype alleles with frequency >1% in the training set.

Results

Fitting haplotype alleles rather than SNPs had the largest improvement in prediction accuracy and reduction of bias for Jerseys, but showed little improvement for Holstein Friesians. The largest improvement in accuracy from fitting haplotypes was 7.7±1.7% when using recombination-based haplotypes for Liveweight in Jerseys. Haplotypes based on multi-locus LD had the lowest prediction accuracy of all methods evaluated. Fitting only haplotype alleles with >1% frequency in the training population typically had negligible impact on prediction accuracy or bias compared to fitting all haplotype alleles.

Conclusions

Haplotype blocks based on pairwise LD performed poorly compared to other methods. Haplotype blocks generated using information on recombination events in the population provided the highest accuracy overall and are recommended for use, provided there are enough parent-offspring pairs to accurately identify recombination events. Fixed length haplotype blocks or blocks that reduce the number of haplotype alleles to be fitted may be suitable to define haplotypes for datasets where recombination events cannot be accurately identified.

3.2 Background

The rate of genetic improvement in a population is limited by the ability to accurately predict the genetic merit of selection candidates (Lush, 1937). Genomic selection using SNP markers is routinely used in dairy cattle breeding programs across the world (Hayes et al., 2013), because of increased prediction accuracy at a young age compared to parental average (VanRaden, 2008). A haplotype block (haploblock) defines a region of the genome, comprising a set of neighboring genetic markers (i.e. SNPs) whose alleles are likely to be inherited together. A haplotype allele is a combination of SNP alleles that are present in a haploblock. When haploblocks are appropriately defined, genomic prediction models that fit haplotype alleles have been shown to improve prediction accuracy over models that fit SNPs using both real (Cuyabano et al., 2015a; Hess et al., 2016) and simulated (Villumsen et al., 2009; Sun et al., 2016) data.

Haploblocks that are the same size throughout the genome have been defined in previous studies, and we refer to these as fixed-length haploblocks. The size of such haploblocks has been defined based on the number of SNPs (Hayes et al., 2007; Calus et al., 2009; Villumsen et al., 2009) or physical map distance (megabases) (Hess et al., 2016; Sun et al., 2016). Recombination rates across the cattle genome are non-random, characterized by areas where recombination occurs more or less frequently, termed recombination hotspots and coldspots, respectively (Sandor et al., 2012; Weng et al., 2014). This variation in recombination rate, along with selective sweeps caused by generations of selection for production traits (Maynard-Smith and Haigh, 1974; Mignon-Grasteau et al., 2005), has resulted in different levels of LD in different genomic regions (Hayes et al., 2009). Haploblocks that vary in length across the genome based on recombination or LD patterns (termed variable-length haploblocks) may result in higher genomic prediction accuracy than fixed-length haploblocks. Rinaldo et al. (2005) described three

variable-length approaches to assign SNPs to haploblocks that have been used in human haplotype studies: 1) minimizing haplotype diversity within a block; 2) location of recombination hotspots; and 3) the level of linkage disequilibrium between loci. Haploblocks generated from these variable-length methods are expected to be population-specific because of different recombination events and LD patterns in different populations (Amaral et al., 2008; de Roos et al., 2008), whereas the fixed-length haploblocks described above will be the same across populations for a given SNP panel.

In Bayesian linear regression models, the effects of rare alleles are shrunk towards zero more so than the effects of common alleles (Gianola, 2013), and it is common practice to remove SNPs with low minor allele frequency (MAF) prior to performing genomic prediction (VanRaden et al., 2009; Harris and Johnson, 2010; Hayes et al., 2010). Hess et al. (2016) showed that removing haplotype alleles present at <10% frequency in the training population had a similar prediction accuracy as removing haplotype alleles at <1% frequency when fixed-length haploblocks of 125 kb or 250 kb were used. However, Hess et al. (2016) did not evaluate prediction accuracy when fitting all haplotype alleles because the number of haplotype alleles was very large, particularly as haploblock size increased (e.g. a haploblock of 20 SNPs has over a million possible haplotype alleles). New haplotype alleles can be generated in a population if there is a recombination event within a haploblock, while recombination events that occur between adjacent haploblocks will not create new haplotype alleles. Assuming the same number of haploblocks are generated, the variable-length approaches described by Rinaldo et al. (2005) are expected to have fewer recombinations within haploblocks and more recombinations between haploblocks than fixed-length haploblocks. Therefore, there are expected to be fewer haplotype alleles from variable-length haploblocks than fixed-length haploblocks, making it

more feasible to run genomic prediction models with all haplotype alleles when variable-length haploblocks are used.

The objectives of this paper were to: 1) compare haploblocks generated from different haploblock methods; 2) evaluate the performance of genomic prediction when fitting covariates for haplotype alleles rather than SNPs when using different haploblock methods; and 3) assess the impact of fitting only haplotype alleles with frequency >1% in the training population on performance of genomic prediction.

3.3 Methods

Phenotype Data

Livestock Improvement Corporation (LIC) provided first lactation yield deviations (YD) (Van Raden and Wiggans, 1991) for milk fat yield (Fat; h^2=0.28), liveweight (Lwt; h^2=0.30) and somatic cell score (SCS; h^2=0.15) (LIC, 2009). Individuals with outlier records or in outlier contemporary groups were filtered out, as in Hess et al. (2016). The training data set contained all genotyped females with YD that were born before 1 June 2008 and the validation data contained later-born genotyped females with YD. The numbers of animals in the training and validation sets for each trait are shown in Table 3.1.

Table 3.1: Numbers of training and validation cows used for genomic prediction

Breed[1]	Fat[2]		Lwt[2]		SCS[2]	
	Training	Validation	Training	Validation	Training	Validation
HF	9,072	3,354	3,908	1,464	9,094	3,358
J	5,067	5,854	2,667	2,331	5,071	5,860
KX	10,684	6,125	6,708	2,436	10,699	6,140
Pooled	24,823[3]	15,333	13,283[3]	6,231	24,864[3]	15,358

1) HF = predominantly (>7/8) Holstein Friesian; J = predominantly (>7/8) Jersey; KX = admixed KiwiCross
2) Yield Deviation: Fat = Milk Fat Yield; Lwt = Liveweight; SCS = Somatic Cell Score
3) Training was performed using pooled data across the three breed classes

Haplotype Construction

Genotype information was collected on either v1 or v2 Illumina BovineSNP50 Beadchips (Matukumalli et al., 2009) or the Illumina BovineHD Beadchip (Matukumalli et al., 2011) for 58,369 dairy cattle born between 1960 and 2012 (females = 46,614; males = 11,755). Animals were phased with LINKPHASE3 and DAGPHASE, as described in the Supplemental Material of Druet and Georges (2015) and in Hess et al. (2016), with 37,740 autosomal SNPs remaining according to the UMD 3.1 map of the Bos taurus genome (Genbank accession: DAAA00000000.2).

Fixed-Length

The fixed-length (FixedLength) haploblocks used in this study were non-overlapping 250 kb regions because this haploblock length had previously been shown to result in the highest accuracy when using the same animals and traits (Hess et al., 2016). That method produced 9,676 unique haploblocks across the genome. All cut-offs for variable-length methods, described below, were selected to produce a similar number of haploblocks in our population as this fixed-length method to facilitate comparison between the different haploblock methods.

Pairwise Linkage Disequilibrium

Appendix A describes the algorithm used to generate haploblocks based on pairwise LD (PairwiseLD) as an example. Pairwise LD was calculated using all individuals that were in the combined training and validation set for any of the three traits (N = 40,065). The LD measurement D' is commonly used to define haploblocks (Jeffreys et al., 2005; Khatkar et al., 2007; Kim and Kirkpatrick, 2009) and fitting haplotype alleles generated using this measure has been shown to improve genomic prediction accuracy in dairy cattle compared to fitting SNPs

(Cuyabano et al., 2014). Using D', haploblocks were constructed using the following iterative joining algorithm:

$$D = Pr(A_1B_1) - Pr(A_1)Pr(B_1) \quad [3.1]$$

$$D' = \begin{cases} \frac{D}{Min\left(Pr(A_1)Pr(B_1), Pr(A_2)Pr(B_2)\right)}, if\ D < 0 \\ \frac{D}{Min\left(Pr(A_1)Pr(B_2), Pr(A_2)Pr(B_1)\right)}, if\ D > 0 \end{cases} \quad [3.2]$$

where A_1 and A_2 are the two alleles at SNP locus A, while B_1 and B_2 are the alleles at SNP locus B, and A_1B_1 refers to a haplotype containing allele 1 at both SNPs (Lewontin, 1964). At the start of this algorithm each haploblock contained only one SNP. The minimum D' (measurement) between each SNP in one haploblock (i.e. SNP A) and each SNP in an immediately adjacent haploblock (i.e. SNP B) was calculated and the neighboring haploblocks that maximized this measurement (joining criteria) were joined each iteration until there were the same number of haploblocks as in the 250 kb fixed-length method. The resulting genome-wide cutoff for D' was 0.1167.

Multi-Locus Linkage Disequilibrium

Multi-locus LD (MultiLocusLD) haploblocks were generated using the same algorithm as the pairwise LD method (Appendix A) but haploblocks were joined using the p-value obtained from a G-test (Sokal and Rohlf, 1995), which is similar to a Chi-square test, with the null hypothesis that the haplotype alleles segregate independently between the two neighboring haploblocks. Neighboring haploblocks with the smallest p-value were joined each iteration (joining criteria).

A table of counts was generated for two adjacent haploblocks, with the same number of rows as the number of unique haplotype alleles at the first haploblock and the same number of

columns as the number of unique haplotype alleles at the second haploblock. The table is filled based on the phased haplotypes. The G-statistic is:

$$G = 2\sum_{i} O_i \cdot ln\left(\frac{O_i}{E_i}\right) \quad [3.3]$$

whereby O_i is the observed count of cell i from the table of counts, E_i is the expected count of cell i under the assumption of independence, and the i's are each of the cells with an observed count greater than zero. The p-value was obtained by dividing the G-statistic by 1,000 and comparing this to a chi-square distribution with $(r-1)\times(c-1)$ degrees of freedom, where r is the number of observed unique haplotype alleles in one haploblock, and c is the number of observed unique haplotype alleles in the adjacent haploblock. It was necessary to divide the G-statistic by 1000 because this dataset was so large that all p-values were zero when using the unadjusted G-statistic. This adjustment can be likened to a sample that is $1/1000^{th}$ of the size of this data set (i.e. 40 individuals, each with 2 haplotypes), with all haplotype alleles present in the same proportions as in the full data set. The genome-wide cutoff for the p-value to obtain a similar number of haploblocks as the 250 kb fixed-length haploblocks was 0.031 after dividing the G-statistic by 1000.

Recombination

The recombination method (Recombination) to assign SNPs to haploblocks used 36,166 parent-offspring pairs from the full dataset of 58,369 individuals to identify recombination events from phased haplotypes. See Appendix B for a detailed description of the method. A complete linkage agglomerative hierarchical clustering method (hclust) was used in R to cluster SNPs into haploblocks (RCoreTeam, 2014). The distance matrix used for the clustering

contained the total number of recombinations between each pair of SNPs and was calculated separately for each chromosome. For example, the number of recombinations between SNPs A and C, separated by SNP B, was the sum of the number of recombinations between SNPs A and B, and SNPs B and C.. A cut-off of 65 recombinations per haploblock resulted in a similar number of haploblocks as the 250 kb fixed-length haploblocks.

Haplotype Alleles

The haplotype allele method (HapAlleles) to assign SNPs to haploblocks aimed to reduce the number of haplotype alleles fit in genomic prediction models. This method was as described in Appendix A, but the measurement was the number of haplotype alleles that would be produced if the two haploblocks were joined and haploblocks that generated the fewest haplotype alleles were joined (joining criteria).

Similarity of Haploblock Methods

The similarity coefficient between haploblock methods was calculated as described in Torres et al. (2009). The similarity of two methods can range from zero (no similarity) and one (complete concordance) and is calculated by:

$$Sim(M_1, M_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{\left|H_i \cap G_j\right|}{\left|H_i \cup G_j\right|} \Big/ max(n_1, n_2) \quad [3.4]$$

where M_1 and M_2 are the two haploblock methods that are being compared, n_1 and n_2 are the number of haploblocks in each method, H_i is haploblock i from method 1, G_j is haploblock j from method 2, $\left|H_i \cap G_j\right|$ is the number of SNPs that are in both H_i and G_j and $\left|H_i \cup G_j\right|$ is the number of SNPs that are in either H_i or G_j, or both.

Genomic Prediction

Genomic Prediction fitting covariates for either SNPs or haplotype alleles was run in GenSel v4.73R (Fernando and Garrick, 2009). Posterior estimates of covariate effects were obtained from a single Markov chain Monte Carlo (MCMC) chain of length 41,000, including 1,000 iterations of burn-in samples that were discarded. Each haploblock model was run by fitting all haplotype alleles that were present in at least five copies across all 40,065 training and validation animals, referred to as "All" haplotype alleles for the remainder of this paper. Haplotype alleles present in less than five copies were removed because this allele frequency is so low that they will likely be shrunk to zero in genomic prediction models, and because their removal reduced the dimensionality of the haplotype matrices. Genomic prediction was also performed after removing covariates for haplotype alleles if they were present at <1% frequency in the training population for that trait. The latter can be considered analogous to filtering SNPs based on MAF.

BayesA

The model for BayesA (Meuwissen et al., 2001) is:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Xh} + \sum_{j=1}^{k} \boldsymbol{z}_j \alpha_j + \boldsymbol{e} \qquad [3.5]$$

where **y** is an N × 1 vector of YD, μ is the intercept, **X** is an incidence matrix of pairwise heterosis fractions between Holstein (H), Friesian (F), Jersey (J) and Red (R) breeds, **h** is a vector of 6 heterosis effects, k is the number of covariates for SNPs (SNP model) or haplotype alleles (haplotype model), $\mathbf{z}_j$ is an N × 1 vector of genotypes (0/1/2) at SNP j (SNP model) or haplotype allele count (0/1/2) at haplotype allele j (haplotype model), α_j is the additive effect of that SNP or haplotype allele, and **e** is an N × 1 vector of identically and independently distributed

residual effects with a mean of zero and variance σ_e^2. The prior for σ_e^2 is a scaled inverse chi-square distribution with scale parameter S_e^2 and ν_e degrees of freedom and marker effects (α_j) are assumed to have identical and independent t-distributions with scale parameter S_α^2 and ν_α degrees of freedom.

Evaluation of Prediction Models

The training set included all breeds pooled (HF, J and KX animals) but performance was evaluated using the validation set for each breed separately. The Direct Genomic Value (DGV) for an individual was calculated as:

$$\widehat{\boldsymbol{DGV}} = \boldsymbol{X}\hat{\boldsymbol{h}} + \sum_{j=1}^{k} \boldsymbol{z}_j \hat{\alpha}_j \qquad [3.6]$$

Fixed heterosis effects ($\boldsymbol{X}\hat{\boldsymbol{h}}$) were added back in because heterosis contributes to the genetic merit of an individual. Model performance was evaluated based on accuracy, calculated as the correlation between YD and DGV, and bias, which is the deviation of the regression coefficient of YD on DGV from 1.

Bootstrap Samples

Bootstrap samples of validation animals were taken to evaluate the error in estimates of accuracy and bias that were due to the validation sample of animals. A bootstrap sample the size of the validation set for that breed was obtained by sampling validation animals for that breed with replacement. Ten thousand bootstrap samples were taken for each breed and the same sample of animals was evaluated for all models to allow paired t-tests to be performed. Accuracy and bias were calculated for each sample and the mean and standard error was taken across all bootstrap samples. Significance was determined based on a p-value threshold of 0.05, with one-

sided t-tests performed to test the improvement of accuracy of haplotype models over the SNP model and to test the improvement of accuracy of variable-length models over the fixed-length models; all other t-tests performed were two-sided.

3.4 Results

Haploblock Method Comparisons

The thresholds used for generating haploblocks in this study successfully constrained the number of blocks to be similar across all methods (Table 3.2). The number of haplotype alleles differed between the methods evaluated, with the MultiLocusLD method generating the most haplotype alleles, followed by the FixedLength method. As intended, the HapAlleles method generated the fewest haplotype alleles of all methods. The FixedLength, MultiLocusLD and HapAlleles methods had a maximum haploblock length of 10 or 11 SNPs, while the PairwiseLD and Recombination methods resulted in haploblocks approximately double this length. Although there was a difference in the maximum number of SNPs per haploblock, the median number of SNPs per haploblock was very similar across methods. The mean number of SNPs per haploblock was 4 for all methods.

Table 3.2: Haploblock Structure from Different Methods

Method[1]	Number Blocks[2]	Number Alleles[3]	SNPs per Haploblock				
			Min	1st Quart.	Median	3rd Quart.	Max
FixedLength	9,676	106,102	1	3	4	5	10
PairwiseLD	9,676	102,892	1	2	3	5	19
MultiLocusLD	9,673	128,174	1	2	4	6	11
Recombination	9,669	101,055	1	3	4	5	20
HapAlleles	9,670	90,916	2	4	4	4	11

1) Haploblock Methods: FixedLength = 250 kb Fixed-length; PairwiseLD = Pairwise LD based on D'; MultiLocusLD = Multi-Locus LD based on G-test; Recombination = based on recombination events; Alleles = Reducing the number of haplotype alleles
2) Number of haploblocks across the genome
3) Number of haplotype alleles across the genome in the data set containing all animals in training and validation sets for all traits

The two most similar haploblock methods in this study were the Recombination and HapAlleles methods (Table 3.3). The two methods that had the lowest similarity were the PairwiseLD and MultiLocusLD methods, indicating that the multi-locus approach is utilizing different information than the pairwise method for evaluating LD between two haploblocks. Both LD methods had low similarity with other methods.

Table 3.3: Similarity between Haploblock Methods

Method	FixedLength	PairwiseLD	MultiLocusLD	Recombination	HapAlleles
FixedLength	1.000	0.667	0.653	0.705	0.704
PairwiseLD	0.667	1.000	0.642	0.680	0.704
MultiLocusLD	0.653	0.642	1.000	0.648	0.661
Recombination	0.705	0.680	0.648	1.000	0.708
HapAlleles	0.704	0.704	0.661	0.708	1.000

The concordance of the haploblock methods around QTL is also of interest because methods that can more accurately capture QTL effects will likely have higher prediction accuracy. The haploblocks around two major QTL on BTA 14: DGAT1 (Grisart et al., 2002) and PLAG1 (Karim et al., 2011) are displayed in Figure 3.1. All haploblock methods had a similar number of haploblocks around DGAT1, except the Recombination method, which had double the others. There was more variation in the number and length of haploblocks around PLAG1 between methods. Although the PairwiseLD and MultiLocusLD methods had the least similar clusters across the genome (Table 3.3), haploblocks were identical around DGAT1 and one of the PairwiseLD haploblocks around PLAG1 was split in two for the MultiLocusLD method (Figure 3.1). The two LD-based methods resulted in haploblocks that spanned each QTL, as did the HapAlleles and FixedLength methods for DGAT1, while for all other cases the SNPs around the QTL were split into up- or down-stream haploblocks.

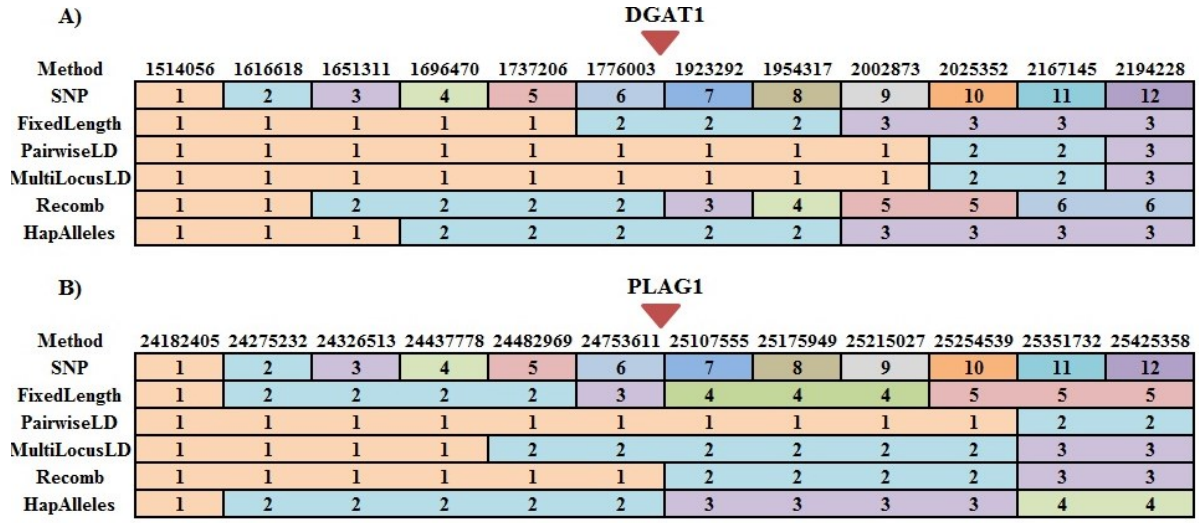


Fig. 3.1: Haploblock Structure around DGAT1 and PLAG1 on BTA 14
Haploblock structure of the 12 SNPs around DGAT1 (A) and PLAG1 (B) differed between the five haploblock methods. Distinct haploblocks for a given method are designated by a different color and number. The position of each SNP along BTA 14 in base pairs is provided at the top of each figure, along with the position of DGAT1 or PLAG1.

Haplotype Model Performance

This section presents accuracy and bias for the models that fit all haplotype alleles, while the next section will cover the impact of fitting only alleles that were present at >1% frequency in the training population. Prediction accuracies for the SNP model can be found in Table 3.4; accuracies for the haplotype models (Table 3.5) are reported as the percent improvement over the SNP model. Prediction bias is reported in Table 3.6.

Table 3.4: Correlation between Yield Deviation and Direct Genomic Value (±SE) from Genomic Prediction fitting SNP Covariates

Trait	Holstein Friesian	Jersey	KiwiCross
Milk Fat Yield	0.348 ± 0.016	0.273 ± 0.013	0.299 ± 0.012
Liveweight	0.398 ± 0.022	0.339 ± 0.019	0.436 ± 0.016
Somatic Cell Score	0.190 ± 0.016	0.147 ± 0.013	0.161 ± 0.013

Prediction Accuracy

The accuracy of haplotype models was found to depend on trait, breed, and haploblock method (Table 3.5). There was no significant improvement in accuracy from fitting a haplotype

model over a SNP model (Table 3.4) for Holstein Friesians for any trait, except when fitting FixedLength haplotype alleles for Fat (Table 3.5). Fitting haplotype alleles improved prediction accuracy in Jerseys and KiwiCross, with higher accuracy for Lwt and SCS for Jerseys, and Fat and Lwt for KiwiCross.

Table 3.5: Percent Change in Genomic Prediction Accuracy from Fitting Haplotype Alleles Rather than SNPs (±SE)

Method[1]	Trait[2]	Holstein Friesian		Jersey		KiwiCross	
		All	1%	All	1%	All	1%
Fixed	Fat	2.3 ± 1.1*	2.1 ± 1.1*	-0.1 ± 1.2	-0.1 ± 1.2	2.4 ± 1.0**	2.3 ± 1.0*
	Lwt	1.9 ± 1.6	2.1 ± 1.5	3.5 ± 1.8*	3.5 ± 1.8*	1.8 ± 1.2	2.3 ± 1.1*
	SCS	0.0 ± 2.3	-0.1 ± 2.3	5.5 ± 2.1**	5.5 ± 2.1**	0.0 ± 2.0	0.1 ± 2.0
PLD	Fat	1.7 ± 1.3	1.5 ± 1.3	-0.4 ± 1.5	-0.4 ± 1.5	1.8 ± 1.2	1.6 ± 1.2
	Lwt	-1.5 ± 1.8	-1.8 ± 1.8	3.6 ± 1.8*	3.3 ± 1.8*	0.0 ± 1.3	-0.2 ± 1.3
	SCS	0.6 ± 2.2	0.5 ± 2.2	2.2 ± 2.1	2.4 ± 2.1	0.9 ± 2.0	0.7 ± 2.0
MLD	Fat	-3.2 ± 1.9	-0.7 ± 1.3	-11.0 ± 2.3	-0.1 ± 1.6	-1.8 ± 2.0	2.8 ± 1.3*
	Lwt	-2.9 ± 2.7	2.0 ± 1.8	1.2 ± 3.0	6.4 ± 2.0**	4.8 ± 2.0**	2.4 ± 1.3*
	SCS	-7.8 ± 3.7	0.3 ± 2.5	-6.6 ± 3.8	3.3 ± 2.3	-10.2 ± 3.4	-0.1 ± 2.2
Rec	Fat	0.2 ± 1.0	0.0 ± 1.0	1.1 ± 1.1	1.0 ± 1.1	0.9 ± 0.9	0.8 ± 0.9
	Lwt	1.2 ± 1.5	0.9 ± 1.5	7.7 ± 1.7**	7.5 ± 1.7**	2.2 ± 1.1*	2.0 ± 1.1*
	SCS	-0.5 ± 2.2	-0.4 ± 2.2	6.9 ± 2.0**	6.8 ± 2.0**	0.1 ± 1.9	0.0 ± 1.9
HA	Fat	1.3 ± 1.0	1.3 ± 1.0	-0.3 ± 1.2	-0.4 ± 1.2	2.5 ± 0.9**	2.3 ± 0.9**
	Lwt	1.9 ± 1.5	1.7 ± 1.6	4.8 ± 1.6**	5.1 ± 1.6**	2.9 ± 1.0**	2.5 ± 1.1*
	SCS	-1.7 ± 2.2	-1.8 ± 2.2	5.6 ± 2.0**	5.7 ± 2.0**	2.1 ± 1.8	2.1 ± 1.8

1) Fixed = Fixed Length; PLD = Pairwise LD; MLD = MultiLocusLD; Rec = Recombination; HA = HapAllele
2) Fat = Milk Fat Yield; Lwt = Liveweight; SCS = Somatic Cell Score
* Higher accuracy than the SNP model ($P < 0.05$)
** Higher accuracy than the SNP model ($P < 0.01$)

Haploblock methods based on measures of LD (PairwiseLD or MultiLocusLD) tended to result in lower prediction accuracy than the other haploblock methods investigated (Table 3.5). When all haplotype alleles were fitted, the MultiLocusLD method frequently had lower prediction accuracy than the SNP model. The Recombination method had the highest accuracy for all traits in Jerseys and showed no significant drop in accuracy in HF or KX. The HapAlleles method had higher accuracies than the Recombination method for KX animals. Similar to the

Recombination method, the FixedLength method either improved or maintained accuracy for all traits and breeds, however accuracies of the Recombination and HapAlleles methods were always higher than the accuracy for the FixedLength method, except for SCS in HF.

Prediction Bias

With the exception of MultiLocusLD, HF and KX predictions were not significantly biased, except for SCS in KX, and there were no reductions in bias compared to the SNP model for either breed. Typically, predictions for Jerseys were biased upwards, except for Fat when all MultiLocusLD haplotype alleles were fitted. The MultiLocusLD methods were almost always biased downwards and significantly different from 1 (Table 3.6). Thus, these two sources of bias (breed and haploblock method) likely cancelled each other out, resulting in a prediction bias of zero for Jersey Fat DGV. The Recombination method decreased bias compared to the SNP model for all traits in Jerseys, as did the HapAlleles method for all traits except Fat.

Filtering Haplotype Alleles

Number of Haplotype Alleles Fitted

Although the MultiLocusLD method produced the most haplotype alleles, many haplotype alleles were at low frequency in the population, so when haplotype alleles with <1% frequency in the training data set were removed the number of haplotype alleles was similar to other haploblock methods (Table 3.7). After filtering at a 1% threshold, the PairwiseLD method had the fewest haplotype alleles, rather than the HapAlleles method.

Table 3.6: Prediction Bias (±SE) of SNP and Haplotype Models

Method[1]	Trait[2]	Holstein Friesian		Jersey		KiwiCross	
		All	1%	All	1%	All	1%
SNP	Fat	-0.04±0.05		**0.16±0.04**		-0.01±0.04	
	Lwt	0.03±0.06		**0.21±0.05**		0.00±0.04	
	SCS	0.05±0.08		**0.23±0.07**		**0.17±0.07**	
Fixed	Fat	-0.05±0.05	-0.05±0.05	**0.16±0.04**	**0.17±0.04**	-0.04±0.04	-0.03±0.04
	Lwt	0.01±0.06	0.01±0.06	**0.17±0.05****	**0.18±0.05***	-0.01±0.04	-0.02±0.04
	SCS	0.05±0.08	0.05±0.08	**0.18±0.07****	**0.18±0.07****	**0.16±0.07**	**0.16±0.07**
PLD	Fat	-0.04±0.05	-0.04±0.05	**0.17±0.04**	**0.18±0.04**	0.00±0.04	0.00±0.04
	Lwt	0.02±0.06	0.02±0.06	**0.16±0.05****	**0.17±0.05****	-0.02±0.04	-0.02±0.04
	SCS	0.04±0.08	0.04±0.08	**0.21±0.07**	**0.21±0.07**	**0.16±0.07**	**0.16±0.07**
MLD	Fat	**-0.32±0.07****	-0.02±0.05	0.00±0.05**	**0.18±0.04**	**-0.29±0.05****	-0.02±0.04
	Lwt	**-0.23±0.08**	0.02±0.06	**-0.16±0.07**	**0.16±0.05****	**-0.37±0.05****	-0.01±0.04
	SCS	**-0.51±0.14***	0.04±0.08	**-0.27±0.13**	**0.19±0.07***	**-0.33±0.12**	**0.16±0.07**
Rec	Fat	-0.04±0.05	-0.03±0.05	**0.14±0.04***	**0.15±0.04**	-0.02±0.04	-0.02±0.04
	Lwt	0.01±0.06	0.02±0.06	**0.14±0.05****	**0.15±0.05****	-0.02±0.04	-0.02±0.04
	SCS	0.05±0.08	0.05±0.08	**0.17±0.07****	**0.18±0.07****	**0.17±0.07**	**0.17±0.07**
HA	Fat	-0.05±0.05	-0.05±0.05	**0.16±0.04**	**0.16±0.04**	-0.04±0.04	-0.04±0.04
	Lwt	0.00±0.06	0.01±0.06	**0.16±0.05****	**0.16±0.05****	-0.03±0.04	-0.03±0.04
	SCS	0.06±0.08	0.06±0.08	**0.18±0.07****	**0.18±0.07****	**0.15±0.07**	**0.15±0.07**

1) Fixed = Fixed Length; PLD = Pairwise LD; MLD = MultiLocusLD; Rec = Recombination; HA = HapAllele
2) Fat = Milk Fat Yield; Lwt = Liveweight; SCS = Somatic Cell Score
* More biased than the SNP model ($P < 0.05$)
** More biased than the SNP model ($P < 0.01$)
Bold Bias is significantly different from zero

Table 3.7: Number of SNP or Haplotype Allele Covariates in the Model

Method	Milk Fat Yield		Liveweight		Somatic Cell Score	
	All	1%	All	1%	All	1%
SNP	37,226		37,356		37,229	
FixedLength	106,022	64,724	105,351	64,634	106,022	64,730
PairwiseLD	102,784	60,177	102,038	60,076	102,784	60,181
MultiLocusLD	128,013	67,929	127,083	67,878	128,014	67,939
Recombination	100,957	62,458	100,289	62,377	100,957	62,459
HapAlleles	90,847	63,251	90,394	63,166	90,847	63,245

Prediction Accuracy and Bias

Filtering haplotype alleles based on frequency had a similar impact on accuracy and bias in each of the breeds (Table 3.5). In some cases, prediction accuracy was improved by only

fitting haplotype alleles that were present at >1% frequency in the training population, although there was often no significant difference in accuracy from fitting all haplotype alleles for a given breed and haploblock method. The exception was the MultiLocusLD method, which performed significantly better when the filter was applied, except for Lwt in KX. Drops in accuracy from fitting only haplotype alleles at >1% frequency compared to all haplotype alleles were negligible and not significant; the largest drop was 0.4% for the Alleles method for KX Lwt predictions. Bias estimates were very similar when fitting all haplotype alleles compared to only those with frequency >1% in the training data set, except when the MultiLocusLD method was fitted, in which case removing rare haplotype alleles reduced bias (Table 3.6).

3.5 Discussion

Haploblock Algorithm

Haploblocks constructed using pairwise LD commonly assign SNPs to haploblocks by starting at one end of each chromosome (typically at 0 bp) and moving along the chromosome, adding a SNP into the previous haploblock or assigning it to a new haploblock (Cuyabano et al., 2014). While the computation efficiency of this method is desirable, different haploblocks are obtained depending upon which end of the chromosome the algorithm is started (Figure S3.1); therefore, haploblocks constructed using this approach may be a poor representation of the true LD structure, which can encumber the interpretation of results obtained from these haploblocks. Alternatively, haploblocks could be constructed by joining SNPs starting from the center of the chromosome; however the resulting haploblocks will still depend on the start position of the algorithm.

The algorithm described in our study to assign SNPs to haploblocks iteratively joined two neighboring haploblocks based on a measurement (e.g. the minimum D') and joining criteria (e.g. maximum measurement), as shown in Appendix A. This algorithm is classified as a greedy algorithm: it makes the locally optimal choice each iteration, with the goal of finding the globally optimal haploblocks (Black, 2004). The globally optimal haploblocks for Pairwise LD are the set of haploblocks that maximize D' between SNPs within haploblocks and minimize pairwise LD between SNPs in different haploblocks. Although this algorithm is not guaranteed to find the globally optimal haploblocks (Black, 2004), it is computationally much faster than an exhaustive search across all possible haploblocks. This algorithm will likely also result in haploblocks that are closer to the optimal haploblocks than the method that begins at one end of the chromosome. The MultiLocusLD and HapAlleles haploblocks were generated using the same algorithm but with different measures for joining (minimum p-value from a G-test and minimum number of haplotype alleles within a block, respectively). It is likely that these two methods have different sets of optimal haploblocks than the PairwiseLD method.

Population Structure

Genomic prediction fitting haplotype alleles has improved accuracy in purebred populations as well as in populations with closely-related breeds (Hayes et al., 2007; Calus et al., 2008; Cuyabano et al., 2015a). The improved accuracy that was observed from fitting haplotype alleles rather than SNPs in these populations may be due to higher LD between haplotypes and QTL than SNPs and QTL (Zondervan and Cardon, 2004), an improved ability to accurately capture relationships between individuals in the population (Habier et al., 2007; Ferdosi et al., 2016), or due to capturing close-range epistatic interactions, for example due to interactions

between SNPs in regulatory elements that influence the expression of a gene (Littlejohn et al., 2014), or within clusters of functionally related genes (e.g. in the major histocompatibility complex (Traherne, 2008)).

Genomic prediction using haplotypes in an admixed population has the additional advantage that haplotype alleles may be able to better capture QTL whose effects differ between breed or that segregate in only one breed (Saatchi et al., 2014). If the association between a SNP and phenotype differs between two breeds in an admixed training set, the estimated effect of the SNP will be the mean of the effects in the different breeds, weighted by the proportion of each of the breeds in the training data set. The data set used in our study consisted of Holstein Friesians, Jerseys and KiwiCross. On average, the KiwiCross animals used in our study had 50% Holstein Friesian and 50% Jersey ancestry, but all training data sets contained more Holstein Friesian animals than Jersey. Thus, the estimated SNP effects are weighted more heavily towards the true effect in Holstein Friesian rather than Jerseys. As observed in Tables 3.5 and 3.6, in general the accuracy and bias of Jersey cow predictions was improved considerably more than that of Holstein Friesian predictions, when fitting haplotype alleles rather than SNPs, suggesting that the haplotype alleles may be picking up differences in QTL effects between breeds better than the SNP model.

Fixed-Length Haploblocks

A number of studies have found that genomic prediction fitting covariates for fixed-length haplotypes results in higher prediction accuracy than fitting covariates for SNPs (Villumsen et al., 2009; Boichard et al., 2012; Hess et al., 2016). The accuracy of fixed-length haplotypes provides a benchmark for genomic prediction performance when using more complex

methods to assign SNPs to haploblocks. It was unexpected that the accuracy of genomic prediction fitting fixed-length haploblocks would perform comparably to genomic prediction fitting variable-length haploblocks in many cases (Table 3.5), because the variable-length haploblocks were constructed using information specific to that population, while the fixed-length haploblocks were not. The accuracy of genomic prediction fitting haplotype alleles is dependent on haploblock length (Calus et al., 2009; Villumsen and Janss, 2009) and prediction accuracy decreases rapidly when haploblocks are too long (Hickey et al., 2013; Hess et al., 2016). Each haploblock method used in our study captures slightly different information, so although genomic prediction accuracy was highest for the fixed-length haploblocks when fitting 250 kb haploblocks (Hess et al., 2016), each of the variable-length haploblock methods may reach its highest prediction accuracy when the genome is split into a different number of haploblocks. This may have given the fixed-length haploblock method an advantage over the variable-length methods. Given the results for accuracy and bias in this population, it is recommended that the Recombination and HapAllele methods be tested at a range of cut-off values and genomic prediction accuracy evaluated to assess whether further improvement can be obtained from fitting haplotype alleles rather than SNPs in genomic prediction models.

Another explanation why there is little difference in prediction accuracy when fitting haplotype alleles from fixed-length versus variable-length haploblocks (Table 3.5) is that the three traits studied are all known to be highly polygenic. Although large QTL have been identified for Fat and Lwt (Grisart et al., 2002; Karim et al., 2011), the majority of the genetic variation for these traits is also explained by the polygenic portion (Pryce et al., 2010). Provided the large QTL are accurately captured by adjacent haplotype alleles, the impact of fitting sub-optimal haploblocks around very small QTL is likely to have little impact on prediction

accuracy, unless the sub-optimal haploblocks are fit around enough QTL to have a substantial impact on the DGV.

Recombination Haploblocks

Haploblocks that were defined based on recombination events in the population were found to either maintain or improve prediction accuracy compared to both the SNP and fixed-length haplotype models in all breeds and traits analyzed. Thus, the ability to accurately and precisely identify recombination events is expected to impact prediction accuracy. Many phasing programs, such as LINKPHASE3 (Druet and Georges, 2015), PHASE (Crawford et al., 2004) and SHAPEIT (O'Connell et al., 2014) output information on recombination events. Thus, the means for construction of haploblocks based on recombination events is readily available as a consequence of phasing genotypes, so this method is one of the more computationally efficient methods for assigning SNPs to haploblocks that were evaluated in this study. Following the identification of recombination events, the steps are: 1) generation of the distance matrix and 2) clustering; which are both computationally minor tasks. Only the fixed-length haploblock method generated the haploblock map faster.

The recombination haploblock method is likely more sensitive to differences in the structure of the data set (e.g. degree of relatedness between individuals or data set size) than the other methods evaluated in our study. Accurate identification of recombination events is critical for the success of this method and which depends on both phasing accuracy and the number of parent-offspring pairs in the data set. The data set used in this study was phased using LINKPHASE3 (Druet and Georges, 2015), which initially takes advantage of pedigree information for phasing, then any regions that remain unphased (or partially phased) after

accounting for pedigree information are phased using DAGs from BEAGLE (Browning and Browning, 2009). Most individuals in the data set used in this study have a genotyped sire (and therefore a number of genotyped paternal half-sibs) and some also have a genotyped dam. As a result, phasing accuracy was high. Data sets that include many unrelated individuals are unlikely to have high phasing accuracy (Weng et al., 2014; Ferdosi et al., 2016) and, thus, haploblocks generated using the recombination method will likely have poor accuracy due to the inability to accurately and precisely identify recombination events. Utilizing the HapAlleles method, discussed in further detail below, may by an attractive alternative when the ability to detect recombination events is limited.

The recombination method is expected to be less sensitive to whether the animals in the data set are admixed or purebred than some of the other methods evaluated in this study. Although recombination events have been found to be breed-specific (Weng et al., 2014), in an admixed-breed data set it will be easier to identify recombination hotspots in the breed (or breeds) that is more common – i.e. Holstein Friesians. This does not appear to be an issue in our data set because Jerseys – the less common breed – showed the greatest improvement in prediction accuracy over the SNP model of all breeds for the Recombination method (Table 3.5).

The ability to identify the precise location in which an effective recombination event has occurred heavily depends on the density of the SNP panel used. In most cases in our population we were able to identify which two SNPs the recombination event occurred between in a parent-offspring pair. Sometimes it was only possible to narrow the recombination event down to a region of a few SNPs if the parent was homozygous for a series of SNPs in that region. Increasing SNP density will allow for improved precision to identify where recombination events occur because the SNPs on a higher-density panel are closer together. An individual may be

heterozygous for SNPs on the higher-density panel in the region that appears homozygous on the lower-density panel, narrowing the region where the recombination could have occurred. The improved precision of identification of recombination events may lead to improved genomic prediction accuracy. Conceptually, sequence level genotyping would provide the highest level of precision. Sequencing is becoming more affordable, thus haplotypes constructed based on sequence may be a feasible approach in the future for improving accuracy, despite the limited improvement in accuracy observed in SNP-based models using sequence or high density SNP genotype information (Su et al., 2012; Erbe et al., 2014; van Binsbergen et al., 2015; Heidaritabar et al., 2016).

Although the recombination method performed well in terms of prediction accuracy across breeds and traits, there is still room for further refinement of this method. The approach taken in this study was chosen to make the methods comparable by constraining the number of haploblocks to those observed using the 250 kb fixed length method. This approach likely did not capture the optimal number of haploblocks, despite the observed improvements in accuracy compared to the SNP methods and, in some cases, the fixed-length method. It is likely that prediction accuracy could be further increased by evaluating different cutoff thresholds and thereby increasing or decreasing the number of haploblocks across the genome. Incorporating a weighting system for the likely prevalence of the newly-created haplotype allele within the population (e.g. a recombination event that is identified between a parent and a widely used sire may be weighted more heavily than a recombination event that is identified between parent and a dam with few offspring) may also improve prediction accuracy.

HapAlleles Haploblocks

The HapAlleles method is an example of the first group of haploblock methods described by Rinaldo et al. (2005): minimizing the haplotype diversity within a block. In our study, haplotype diversity was defined as the number of unique haplotype alleles across the genome and the optimal haploblocks for this method was the set of haploblocks that reduced diversity within a block and maximized diversity between haploblocks.

The HapAlleles method tended to have slightly higher accuracy than the Recombination method for Holstein Friesian and KiwiCross cows but lower accuracy than the Recombination method for Jerseys. Therefore, although the gains in accuracy were more moderate for the HapAlleles method than for the Recombination method, the HapAlleles method was one of the best-performing genomic prediction methods. In particular in data sets where it is difficult to accurately identify recombination events, i.e. due to small sample size or not many parent-offspring pairs, the HapAlleles method may outperform the Recombination method. One downside of the HapAlleles method compared to the Recombination method is that it is much more computationally intensive and takes longer to generate the haploblocks.

As SNP density increases, particularly as it approaches sequence density, there will be high LD between more SNPs and clustering SNPs into haploblocks that aim to minimize the number of covariates to be fit in genomic prediction models may improve prediction accuracy because a smaller proportion of haplotype alleles associated with the trait will be rare and shrunk to zero (Gianola, 2013). The HapAllele method may be further improved by weighting the haplotype alleles proportional to their frequency in the population rather than counting the number of unique haplotype alleles. In this approach, haploblocks that generate haplotype alleles

where most of them are common and few of them are rare will be favored, which will improve power to estimate effects for those haplotype alleles.

LD-Based Haploblocks

Our study evaluated two methods for assigning SNPs to haploblocks based on LD information: Pairwise and Multi-Locus. Use of haploblocks based on Pairwise LD improved genomic prediction over fitting SNPs in purebred populations and populations with closely-related breeds (Cuyabano et al., 2014; Cuyabano et al., 2015a; Cuyabano et al., 2015b); however, there is debate as to whether to use the minimum (Cuyabano et al., 2014) or average (Reich et al., 2001) measurement of LD between SNPs to assign SNPs to haploblocks. Therefore, a measurement of multi-locus LD was evaluated to account for correlations between SNPs as a group, rather than pairwise, with the hypothesis that this would improve genomic prediction accuracy. Zhao et al. (2005) found that a standardized Chi-square statistic was a good measure of LD between a multi-allelic marker and a QTL. The G-test is similar to the chi-square test but is less affected by the presence of cells with counts less than five (Sokal and Rohlf, 1995).

Haploblock methods based on LD measurements (PairwiseLD and MultiLocusLD) generally did not perform as well as the other variable-length methods for genomic prediction, particularly when all MultiLocusLD haplotype alleles were fitted (Tables 3.5 and 3.6). The study by Cuyabano et al. (2015a) used 770,000 SNPs rather than the 50,000 SNPs used in our study and their largest haploblock (62 SNPs) would cover approximately the same physical distance as an average-sized haploblock in our study (4 SNPs; Table 3.2). The D' threshold that was needed to obtain the appropriate number of haploblocks in our study was much lower than the threshold of 0.45 used in Cuyabano et al. (2015a). This suggests that the accuracy of genomic prediction

when fitting PairwiseLD haplotype alleles may improve if the genome was split into a greater number of haploblocks; however the majority of haploblocks used by Cuyabano et al. (2015a) would be smaller than the distance between neighboring SNPs at the 50,000 SNP density used in our study.

The population used in the Cuyabano et al. (2015a) study consisted of a number of Nordic dairy cattle breeds with recent common ancestry. In contrast, the New Zealand dairy cattle population used in our study contains genetically distinct breeds that have been extensively crossed in recent generations (LIC and DairyNZ, 2015; Hess et al., 2016). Patterns of linkage disequilibrium are population-specific, and therefore differ between breeds (de Roos et al., 2009). The method that was used to calculate pairwise LD used the measurement D', which is calculated based on the deviation of the frequency of the observed 2-SNP haplotypes compared to what is expected based on the allele frequencies at each of the SNPs under the assumption of independence. If the LD patterns in Holstein Friesians and Jerseys are different in a region, the LD measurements calculated by assuming the two breeds are from the same population may be uninformative as to the true associations between SNPs in those regions in each breed. Therefore, in an admixed-breed population, LD-based haploblock methods may not cluster SNPs into meaningful haploblocks, which may explain the generally poor results for LD-based methods (Tables 3.5 and 3.6).

The MultiLocusLD method could be further improved by using other methods to quantify LD. Our study used p-values from a G-test; however the G-statistic had to be divided by 1000 to obtain p-values that were greater than zero and this method may be sensitive to the number of haplotype alleles in each haploblock (i.e. number of rows or columns in the table of counts). An alternative method that should be explored is the Symmetric Uncertainty Coefficient (Press et al.,

1992), a measurement of association that ranges from zero to one and would not be sensitive to the number of haplotype alleles within each haploblock. However, multi-locus LD methods are very time consuming to compute across the whole genome, particularly when haploblocks become long, and may be more useful when generating haploblocks in a small region of a chromosome.

Computation Time of Haplotype Models

Haplotype models are computationally more intensive than the corresponding SNP model. In addition to the commonly used SNP quality control practices employed prior to their inclusion in a prediction model based on SNPs covariates, for a haplotype model it is also necessary to 1) phase the SNP genotypes; 2) assign SNPs to haploblocks; 3) generate haplotype alleles based on those haploblocks; and 4) perform any filtering on the haplotype alleles. At the SNP density used in our study, haplotype models typically fit more covariates than the corresponding SNP model and therefore genomic prediction takes longer to run. There are multiple ways to reduce the increased computation time associated with haplotype models, as described below.

The computational demands of different methods for assigning SNPs to haploblocks can vary greatly (i.e. the slow MultiLocusLD method vs. the fast FixedLength method); however, this process may not need to be repeated each generation. In most genomic evaluation programs, the training data set grows each generation as additional animals are genotyped and phenotyped however the majority of training animals remain the same. Therefore, LD and recombination patterns in the training population are unlikely to change substantially from one generation to the next (Heifetz et al., 2005) – particularly if the training data set is large. Although from one

generation to the next these patterns may not change substantially, over multiple generations the optimal haploblocks may change and the haploblock map should be updated. The frequency of how often the haploblock map should be updated will likely be dependent on method so the robustness of the haploblock bounds from each of these methods across generations should be evaluated in order to assess how sensitive prediction accuracy is to assuming no change in haploblock structure across generations.

The reduction in the number of haplotype allele covariates that are fit in genomic prediction models will also decrease computation time of haplotype models. This study showed that applying a haplotype allele frequency filter of 1% reduced the number of covariates fit in haplotype models by approximately 40% compared to fitting covariates for all haplotype alleles (Table 3.7) which either improved or maintained prediction accuracy, with substantial improvement for the MultiLocusLD method (Table 3.5). It is important to not set the frequency filter too high, or prediction accuracy and bias will be detrimentally affected. The optimal frequency to reduce dimensionality without reducing accuracy will be dependent on method (Table 3.5) and the length of the haploblocks (Hess et al., 2016). Other approaches to reducing the number of covariates fitted in genomic prediction models are to fit haplotype alleles only in regions of known QTL and SNPs across the remainder of the genome, as in Boichard et al. (2012); or to evaluate the regions that explain more genetic variance under the haplotype model than the SNP model and fit haplotypes only in these regions and SNPs in the remainder of the genome.

3.6 Conclusions

Fitting covariates for haplotype alleles rather than SNPs can improve genomic prediction accuracy and bias in this admixed dairy cattle population. Haploblocks that were generated using a clustering algorithm based on recombination events showed the largest improvements in accuracy of the haploblock methods evaluated, particularly in Jersey cattle. LD-based methods for assigning SNPs to haploblocks did not perform as well as methods based on recombination events or on reducing the number of haplotype alleles, likely because LD is breed- and population-specific and therefore not as informative when calculated using an admixed population. The haploblock method that reduced the number of haplotype alleles across the genome clustered SNPs similar to the recombination-based method and had the highest accuracy for Holstein and KiwiCross breeds and would therefore be a useful method to explore if recombination events cannot be identified accurately. Removing rare haplotype alleles with <1% frequency in the training data set generally maintained or improved genomic prediction accuracy compared to fitting all haplotype alleles. As more individuals in a population are genotyped, it will be possible to improve phasing accuracy and the ability to accurately and precisely identify recombination events, which may lead to further improvement in genomic prediction accuracy when haploblocks are defined based on recombination events.

3.7 Acknowledgements

The authors would like to thank Kathryn Tiplady and Dr. Bevin Harris from Livestock Improvement Corporation for providing the Yield Deviation phenotypes. The authors would also like to thank Dr. Anna Wolc, Dr. Rohan Fernando, Dr. Alicia Carriquiry and Dr. Ken Koehler for their discussions on measurements of multi-locus LD.

3.8 Author Contributions

MH designed the study, ran the analyses, interpreted the results and wrote the manuscript. AH assisted with the study design and interpretation of results. JD contributed to study design and DG supervised the study. All authors critically contributed to the manuscript.

3.9 References

Amaral, A. J., H. J. Megens, R. Crooijmans, H. C. M. Heuven, and M. A. M. Groenen. 2008. Linkage disequilibrium decay and haplotype block structure in the pig. Genetics 179(1):569-579. doi: 10.1534/genetics.107.084277

Black, P. E. 2004. Dictionary of algorithms and data structures. National Institute of Standards and Technology.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Animal Production Science 52(2-3):115-120. (Review) doi: 10.1071/an11119

Browning, B. L., and S. R. Browning. 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. American Journal of Human Genetics 84(2):210-223. doi: 10.1016/j.ajhg.2009.01.005

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178(1):553-561. (Article) doi: 10.1534/genetics.107.080838

Calus, M. P. L., T. H. E. Meuwissen, J. J. Windig, E. F. Knol, C. Schrooten, A. L. J. Vereijken, and R. F. Veerkamp. 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. Genetics, Selection, Evolution 41(11):(15 January 2009). (article)

Crawford, D. C., T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder, D. A. Nickerson, and M. Stephens. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nature Genetics 36(7):700-706. doi: 10.1038/ng1376

Cuyabano, B. C. D., G. Su, G. J. M. Rosa, M. S. Lund, and D. Gianola. 2015a. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. Journal of Dairy Science 98(10):7351-7363. doi: 10.3168/jds.2015-9360

Cuyabano, B. C. D., G. S. Su, and M. S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. Bmc Genomics 15doi: 10.1186/1471-2164-15-1171

Cuyabano, B. C. D., G. S. Su, and M. S. Lund. 2015b. Selection of haplotype variables from a high-density marker map for genomic prediction. Genetics Selection Evolution 47:11. (Article) doi: 10.1186/s12711-015-0143-3

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183(4):1545-1553. doi: 10.1534/genetics.109.104935

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179(3):1503-1512. (Article) doi: 10.1534/genetics.107.084301

Druet, T., and M. Georges. 2015. LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. Bioinformatics 31(10):1677-1679. doi: 10.1093/bioinformatics/btu859

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2014. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 97(10):6622-6622. doi: 10.3168/jds.2014-97-10-6622

Ferdosi, M. H., J. Henshall, and B. Tier. 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genetics Selection Evolution

Fernando, R. L., and D. J. Garrick. 2009. GenSel - User Manual for a Portfolio of Genomic Selection Related Analyses. Second edition. Iowa State University. http://bigs.ansci.iastate.edu/bigsgui/help.html. Accessed 27 June 2016.

Gianola, D. 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194(3):573-596. (Article) doi: 10.1534/genetics.113.151753

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12(2):222-231. (Article) doi: 10.1101/gr.224202

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389-2397. (Article) doi: 10.1534/genetics.107.081190

Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. Journal of Dairy Science 93(3):1243-1252. doi: 10.3168/jds.2009-2619

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92(2):433-443. (Review) doi: 10.3168/jds.2008-1646

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetics Research 89(4):215-220. (Article) doi: 10.1017/s0016672307008865

Hayes, B. J., H. A. Lewin, and M. E. Goddard. 2013. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends in Genetics 29(4):206-214. doi: 10.1016/j.tig.2012.11.009

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. Plos Genetics 6(9)doi: 10.1371/journal.pgen.1001139

Heidaritabar, M., M. P. L. Calus, H. J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, and M. Soller. 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics 171(3):1173-1181. doi: 10.1534/genetics.105.040782

Hess, M., T. Druet, A. Hess, and D. Garrick. 2016. Fixed length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Submitted to GSE*

Hickey, J. M., B. P. Kinghorn, B. Tier, S. A. Clark, J. H. J. van der Werf, and G. Gorjanc. 2013. Genomic evaluations using similarity between haplotypes. Journal of Animal Breeding and Genetics 130(4):259-269. doi: 10.1111/jbg.12020

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers, and P. Donnelly. 2005. Human recombination hot spots hidden in regions of strong marker association. Nature Genetics 37(6):601-606. doi: 10.1038/ng1565

Karim, L., H. Takeda, L. Lin, T. Druet, J. A. C. Arias, D. Baurain, N. Cambisano, S. R. Davis, F. Farnir, B. Grisart, B. L. Harris, M. D. Keehan, M. D. Littlejohn, R. J. Spelman, M. Georges, and W. Coppieters. 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nature Genetics 43(5):405-+. (Article) doi: 10.1038/ng.814

Khatkar, M. S., K. R. Zenger, M. Hobbs, R. J. Hawken, J. A. L. Cavanagh, W. Barris, A. E. McClintock, S. McClintock, P. C. Thomson, B. Tier, F. W. Nicholas, and H. W. Raadsma. 2007. A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein-Friesian cattle. Genetics 176(2):763-772. doi: 10.1534/genetics.106.069369

Kim, E. S., and B. W. Kirkpatrick. 2009. Linkage disequilibrium in the North American Holstein population. Animal Genetics 40(3):279-288. doi: 10.1111/j.1365-2052.2008.01831.x

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics, 49(1), 49-67. Genetics 49(1):49-67.

LIC. 2009. Your Index Your Animal Evaluation System.

LIC and DairyNZ. 2015. New Zealand Dairy Statistics 2014-15, http://www.dairynz.co.nz/media/3136117/new-zealand-dairy-statistics-2014-15.pdf.

Littlejohn, M. D., K. Tiplady, T. Lopdell, T. A. Law, A. Scott, C. Harland, R. Sherlock, K. Henty, V. Obolonkin, K. Lehnert, A. MacGibbon, R. J. Spelman, S. R. Davis, and R. G. Snell. 2014. Expression Variants of the Lipogenic AGPAT6 Gene Affect Diverse Milk Composition Phenotypes in Bos taurus. Plos One 9(1):12. (Article) doi: 10.1371/journal.pone.0085757

Lush, J. L. 1937. Animal breeding plans. Animal breeding plans.:Pp. x + 350. (Book)

Matukumalli, L., S. Schroeder, S. DeNise, T. Sonstegard, C. Lawley, M. Georges, W. Coppieters, K. Gietzen, J. Medrano, and G. Rincon. 2011. Analyzing LD blocks and CNV segments in cattle: novel genomic features identified using the BovineHD BeadChip. San Diego, CA: Illumina Inc

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. Plos One 4(4):13. (Article) doi: 10.1371/journal.pone.0005350

Maynard-Smith, J., and J. Haigh. 1974. Hitch-hiking effect of a favorable gene. Genetics Research 23(1):23-35. (Article) doi: 10.1017/s0016672300014634

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829. (Article)

Mignon-Grasteau, S., A. Boissy, J. Bouix, J. M. Faure, A. D. Fisher, G. N. Hinch, P. Jensen, P. Le Neindre, P. Mormede, P. Prunet, M. Vandeputte, and C. Beaumont. 2005. Genetics of adaptation and domestication in livestock. Livestock Production Science 93(1):3-14. doi: 10.1016/j.livprodsci.2004.11.001

O'Connell, J., D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, and J. Marchini. 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. Plos Genetics 10(4)doi: 10.1371/journal.pgen.1004234

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1992. Numerical Recipes: the Art of Scientific Computing. 3rd ed. Cambridge University Press.

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J. Bowman, K. Savin, M. E. Goddard, and B. J. Hayes. 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. Journal of Dairy Science 93(7):3331-3345. doi: 10.3168/jds.2009-2893

RCoreTeam. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. Nature 411(6834):199-204. doi: 10.1038/35075590

Rinaldo, A., S. A. Bacanu, B. Devlin, V. Sonpar, L. Wasserman, and K. Roeder. 2005. Characterization of multilocus linkage disequilibrium. Genetic Epidemiology 28(3):193-206. doi: 10.1002/gepi.20056

Saatchi, M., R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. Bmc Genomics 15:16. (Article) doi: 10.1186/1471-2164-15-442

Sandor, C., W. B. Li, W. Coppieters, T. Druet, C. Charlier, and M. Georges. 2012. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. Plos Genetics 8(7):13. (Article) doi: 10.1371/journal.pgen.1002854

Sokal, R. R., and F. J. Rohlf. 1995. Biometry : the principles and practice of statistics in biological research. 3rd ed. Freeman, New York.

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. R. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (similar to 54,000) and high-density (similar to 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95(8):4657-4665. (Article) doi: 10.3168/jds.2012-5379

Sun, X., H. Su, P. Boddhireddy, and D. Garrick. 2016. Haplotype-based Genomic Prediction of Breeds Not in Training. In: Plant and Animal Genomes Conference XXIV, San Diego, CA

Torres, G. J., R. B. Basnet, A. H. Sung, S. Mukkamala, and B. M. Ribeiro. 2009. A similarity measure for clustering and its applications. International Journal of Electrical,Computer, and Systems Engineering 3(3):164-170.

Traherne, J. A. 2008. Human MHC architecture and evolution: implications for disease association studies. International Journal of Immunogenetics 35(3):179-192. doi: 10.1111/j.1744-313X.2008.00765.x

van Binsbergen, R., M. P. L. Calus, M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47:13. (Article) doi: 10.1186/s12711-015-0149-x

Van Raden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91(11):4414-4423. (Article) doi: 10.3168/jds.2007-0980

Van Raden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92(1):16-24. doi: http://dx.doi.org/10.3168/jds.2008-1514

Van Raden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal-model information. Journal of Dairy Science 74(8):2737-2746. doi: 10.3168/jds.S0022-0302(91)78453-1

Villumsen, T. M., and L. Janss. 2009. Bayesian genomic selection: the effect of haplotype length and priors. BMC proceedings 3 Suppl 1:S11.

Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics 126(1):3-13. (Article) doi: 10.1111/j.1439-0388.2008.00747.x

Weng, Z. Q., M. Saatchi, R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genetics Selection Evolution 46doi: 10.1186/1297-9686-46-34

Zhao, H., D. Nettleton, M. Soller, and J. C. M. Dekkers. 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genetical Research 86(1):77-87. doi: 10.1017/s001667230500769x

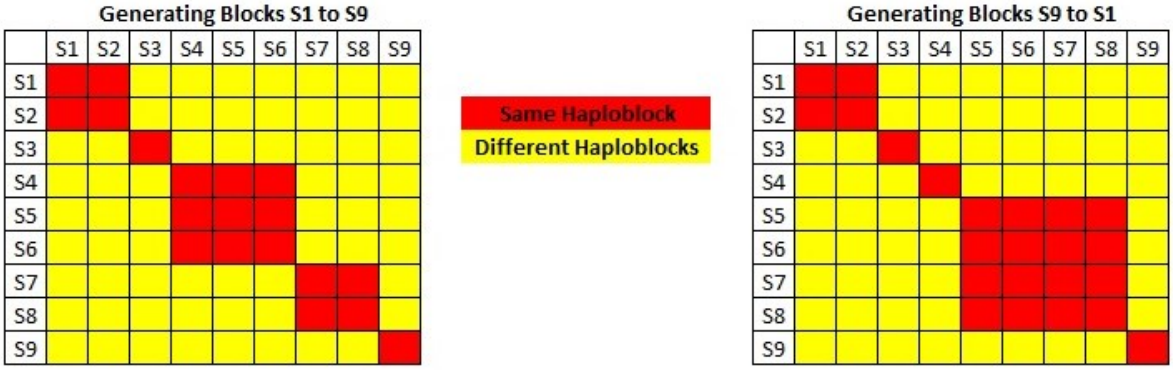
Zondervan, K. T., and L. R. Cardon. 2004. The complex interplay among factors that influence allelic association. Nature Reviews Genetics 5(2):89-U14. (Review) doi: 10.1038/nrg1270

3.10 Figures

Phased Haplotypes for 5 Individuals at 9 SNPs

ID	Strand	S1	S2	S3	S4	S5	S6	S7	S8	S9
1	S	1	1	1	2	2	2	1	1	1
1	D	1	1	2	2	2	2	1	2	1
2	S	1	2	2	1	2	2	1	2	2
2	D	2	1	2	2	1	2	2	1	2
3	S	1	2	1	1	2	2	1	2	1
3	D	1	1	1	1	2	2	2	2	1
4	S	1	1	2	2	2	1	1	1	2
4	D	2	1	2	1	2	2	1	2	2
5	S	1	2	1	2	2	2	2	1	1
5	D	1	1	1	2	2	2	2	1	2

1 = allele 1
2 = allele 2



Haploblocks generated starting from S1 and moving to S9 using the PairwiseLD1 method give different haploblocks than starting at S9 and moving to S1 even though they were generated using the same set of animals. The cutoff for these haploblocks was D' = 0.45.

Additional File S3.1: Defining Haploblocks based on D' from one end of the Chromosome

CHAPTER IV

EVALUATING THE RELIABILITY OF GENOMIC PREDICTION ACCURACY ESTIMATES

Melanie Hayr[1], Andrew Hess[1], Dorian Garrick[1,2], and Rohan Fernando[1]

[1]Iowa State University, Iowa, USA

[2]Massey University, Palmerston North, New Zealand

A paper to be submitted to *PLOS Genetics*

4.1 Abstract

Genomic prediction is a widely used method for prediction of phenotype or genetic merit in plants and livestock, and is increasingly being used for prediction of genetic predisposition to disease in humans. Estimates of marker effects are obtained from a training set of individuals and these estimates are used to predict the genetic merit in a distinct validation set to obtain an estimate of prediction accuracy. We describe a method to obtain the posterior distribution of prediction accuracy by calculating accuracy in each iteration of a Markov chain Monte Carlo chain when performing genomic prediction using a Bayesian model. We show that the posterior distribution of accuracy approximates variation in estimates due to differences in training individuals (n=5000 or 20000), validation individuals (n=1000) and uncertainty in model estimates of effects using a simulated data set of 1 Morgan across 10 chromosomes for a trait

with heritability 0.5 when using a Bayesian GBLUP-like model. The average accuracy within an iteration was not equal to the accuracy of the estimated genetic merit from the model; however when the number of individuals increased compared to the number of markers, the mean of the posterior distribution of accuracy approached the accuracy of the posterior estimates of genetic merit. Our approach for evaluating the reliability of an accuracy estimate is much faster than cross-validation and bootstrapping methods. In contrast to comparing only estimates of prediction accuracy, comparing the posterior distributions of accuracy for two genomic prediction models can enable more informed decisions as to the most appropriate model to use for genomic prediction. Further studies need to investigate whether the posterior distribution of accuracy appropriately captures the uncertainty in genomic prediction accuracy estimates under different scenarios, such as heritabilities and degree of relationships within and between training and validation sets.

4.2 Introduction

The cost of genotyping continues to rapidly decrease (Wetterstrand) and SNP panels are available that can genotype an individual at hundreds of thousands or millions of SNPs in humans (Gray et al., 2000), plants (Unterseer et al., 2014) and livestock (Matukumalli et al., 2009). As a consequence, data sets are available on thousands of individuals genotyped at tens of thousands of markers, which can be used to test associations between genetic variants and phenotypes, a process known as a Genome Wide Association Study (GWAS), or to predict the genetic merit of individuals that do not yet have observed phenotypes, known as genomic prediction (Goddard, 2009). Many traits of interest in humans, plants and livestock are complex, and are likely influenced by many genetic variants (Goddard, 2001). Unlike monogenic or

oligogenic traits (Spichenok et al., 2011), polygenic traits are typically unable to be predicted with high accuracy when few SNPs are used (de los Campos et al., 2013b). Prediction accuracy estimates are usually higher for complex traits when genomic prediction is used rather than prediction from only those SNPs that reach a defined significance threshold in a GWAS (Solberg et al., 2008; Allen et al., 2010).

Most of the early data sets used for genomic prediction contained many more SNPs than individuals; therefore, genomic prediction models regressed phenotypes on all markers simultaneously as random effects (Meuwissen et al., 2001). Those first Bayesian genomic prediction models combined prior assumptions about the distribution of SNP effects with genotype and phenotype information on a set of individuals, termed the training set, to obtain posterior distributions of marker effects for every SNP. The posterior mean SNP estimates can then be used to predict the genetic merit of individuals not in the training set (Meuwissen et al., 2001). A number of Bayesian genomic prediction models with different prior assumptions have since been described (Kizilkaya et al., 2010; Habier et al., 2011; Zeng, 2015) and have been successfully implemented in breeding programs in livestock and plants (de los Campos et al., 2013a). Bayesian genomic prediction methods have also been applied in humans to predict phenotypes of complex traits (de los Campos et al., 2013b).

The accuracy of genomic prediction is typically estimated based on a dataset of genotyped and phenotyped individuals that are split into training and validation sets. The estimate of accuracy is specific to that pair of training and validation sets and different sets will likely give different prediction accuracy estimates (Saatchi et al., 2011). The variation in accuracy estimates between different groups is influenced by the size of the training and validation data sets, the relationship of individuals within and between those data sets (Habier et

al., 2007) and by the trait itself (i.e. heritability and genetic architecture) (Goddard, 2009). Formally quantifying the confidence in the prediction accuracy from a particular data set will serve as a valuable tool for comparing different models or determining the feasibility of implementing genomic prediction (i.e. testing whether the accuracy is expected to exceed a given threshold).

Two methods that are commonly used to evaluate confidence in estimates of prediction accuracy are cross-validation and bootstrapping of the validation set. Cross-validation for genomic prediction is often done via k-means clustering whereby individuals are separated into k groups such that relationships within each group is maximized and relationships between groups are reduced (Saatchi et al., 2011). A training set then comprises k-1 groups and accuracy is estimated in the validation group left out from training; this repeated until an estimate of accuracy is obtained for each group. This method obtains an estimate of accuracy and an indication of confidence in that estimate, however it is time intensive and usually does not appropriately represent relationships between the training and validation data sets that occur in practice. Bootstrapping can be applied to the validation individuals, whereby individuals are sampled with replacement to obtain a bootstrap sample the same size as the original validation data set and prediction accuracy is estimated for this sample. Many bootstrap samples, typically 10,000, are then taken and the prediction accuracy estimate is the average accuracy across all samples and the standard error of this estimate is the standard deviation across all samples (Hess et al., 2016). The bootstrapping method captures variation in the accuracy due to the validation set but does not capture uncertainty in model estimates or a different training data set.

We propose a method to sample the posterior distribution of prediction accuracy using the SNP effects from a single set of Markov chain Monte-Carlo (MCMC) samples. This method

is expected to capture variation in prediction accuracy estimates due to uncertainty in model estimates in training as well as variation due to the individuals included in the training and validation sets. We compared the posterior distribution of accuracy from each analysis to the distribution of prediction accuracy across 200 replicates of training and validation sets from a simulated population.

4.3 Materials and Methods

Simulation

The simulation used in this study is similar to Karaman et al. (2016). Whole-sequence haplotypes from Chromosomes 13 – 22 from the 1,000 Genomes Project (Altshuler et al., 2015) were downloaded from ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Founders used for this simulation consisted of 90 individuals from Great Britain (GBR). A 0.1 cM region from each of the ten downloaded chromosomes was randomly selected for use in the simulation and variants were removed if they were multi-allelic, fixed in the 90 individuals or did not map to a unique position on the chromosome. Founder haplotypes were sampled from the 90 GBR individuals using XSim (Cheng et al., 2015) in Julia to generate a population of size 10,000 (Figure 4.1A). This founder population was randomly mated for 100 non-overlapping generations to obtain a base population of 10,000 individuals. Variants with a minor allele frequency (MAF) < 0.005 in the base population were removed from the analysis; 300 variants were randomly chosen to be QTL and 20,000 to represent markers across the ten simulated chromosomes were randomly chosen as SNPs to be used for genomic prediction. Effects were assigned to QTL based on randomly sampling from a Standard Normal distribution. Appendix C contains the annotated code for the generation of the base population.

Genetic Merit and Phenotypes

The code to generate training and validation populations from the base population (Figure 4.1B) in Scenario 1 (Table 4.1) is provided in Appendix D. Training and validation populations contained 3,000 individuals each generation from random matings between 500 male and 500 female parents. The training population consisted of three generations (Gen 101-103) and individuals from the validation population were all from the same subsequent generation (Gen 104). Individuals from the training and validation populations were randomly selected to obtain data sets of 5,000 and 1,000 individuals, respectively. Scenarios 2 and 3 (Table 4.1) were generated using only the first five simulated chromosomes and for Scenario 3 the number of individuals and parents in Figure 4.1B was multiplied by 4 to generate a population large enough to sample 20,000 training individuals.

Table 4.1: Parameters of the Simulated Data Sets

Scenario	Data Set Size		Chromosomes	Number of SNPs	QTL
	Training	Validation			
1	5,000	1,000	10	20,000	300
2	5,000	1,000	5	10,000	150
3	20,000	1,000	5	10,000	150

The genetic merit of an individual (u_i) was calculated as:

$$u_i = \sum_{j=1}^{n_{QTL}} q_{ij}\alpha_j \quad [4.1]$$

where n_{QTL} is the number of QTL (300), q_{ij} is the 0/1/2 genotype of individual i at QTL j and α_j is the additive effect of that QTL (Appendix C). The phenotype (y_i) for each individual was calculated as:

$$y_i = u_i + e_i \quad [4.2]$$

where e_i represents the non-additive-genetic portion of individual i's phenotype and was drawn from a standard normal distribution scaled to obtain phenotypes with the desired heritability of 0.5.

Bayesian Genomic Prediction

A Bayesian G-BLUP-like model was implemented for genomic prediction that utilized a Single Value Decomposition as outlined in Appendix E.

Sampling the Posterior Distribution of Accuracy

The predicted genetic merit of an individual from the posterior mean marker effect estimates ($\hat{\alpha}$) will be denoted $\hat{u}$ whereas the sampled genetic merit based on the marker effects from a single MCMC sample (α*) will be denoted u*. The accuracy, defined as the correlation between y and $\hat{u}$ can be calculated as:

$$r_{(y,\hat{u})} = \frac{Cov(y,\hat{u})}{\sqrt{Var(y)}\sqrt{Var(\hat{u})}} \quad [4.3]$$

whereas a single sample of accuracy can be calculated as:

$$r_{(y,u^*)} = \frac{Cov(y,u^*)}{\sqrt{Var(y)}\sqrt{Var(u^*)}}. \quad [4.4]$$

The posterior distribution of accuracy is represented by sampling $r_{(y,u^*)}$ for each MCMC sample.

This calculation can also be performed for the accuracy of predicting the true genetic merit (u) in the simulated data set because the true genetic merit is known:

$$r_{(u,\hat{u})} = \frac{Cov(u,\hat{u})}{\sqrt{Var(u)}\sqrt{Var(\hat{u})}} \quad [4.5]$$

whereas a single sample of accuracy can be calculated as:

$$r_{(u,u^*)} = \frac{Cov(u,u^*)}{\sqrt{Var(u)}\sqrt{Var(u^*)}}. \quad [4.6]$$

In a real data set it is not possible to directly calculate $r_{(u,\hat{u})}$ or $r_{(u,u^*)}$ because the true genetic merit of an individual (u) is not observed. It is therefore necessary to approximate the accuracy of predicting true genetic merit. Assuming the covariance between e in the validation data and $\hat{u}$ from the training data is zero:

$$\begin{aligned} Cov(y,\hat{u}) &= Cov(u+e,\hat{u}) \\ &= Cov(u,\hat{u}) + Cov(e,\hat{u}) \\ &= Cov(u,\hat{u}) \end{aligned} \quad [4.7]$$

so $Cov(u,\hat{u})$ can be computed as $Cov(y,\hat{u})$. The variance of u can be estimated by the posterior estimate of genetic variance, so an approximation for the accuracy of predicting the true genetic merit is:

$$\tilde{r}_{(u,\hat{u})} = \frac{Cov(y,\hat{u})}{\sqrt{\widehat{Var(u)}}\sqrt{Var(\hat{u})}} \quad [4.8]$$

where $\widehat{Var(u)}$ is the posterior estimate of genetic variance. This is an equivalent equation to dividing Equation 4.3 by the square root of heritability. It then follows that an approximation for a single MCMC sample is:

$$\tilde{r}_{(u,u^*)} = \frac{Cov(y,u^*)}{\sqrt{\widehat{Var(u)}}\sqrt{Var(u^*)}}. \quad [4.9]$$

Evaluating the Posterior Distribution of Accuracy

The Scenario 1 simulation was replicated 200 times from a single base population and set of QTL effect estimates, whereas Scenario 2 was simulated once and the Scenario 3 simulation

was replicated 160 times. Within each of five randomly selected replicates for Scenario 1, the posterior distributions of accuracy (i.e. distributions of $r_{(y,u^*)}$, $r_{(u,u^*)}$ and $\tilde{r}_{(u,u^*)}$) were compared to the point estimates of the accuracies of the estimated genetic merit ($r_{(y,\hat{u})}$, $r_{(u,\hat{u})}$ and $\tilde{r}_{(u,\hat{u})}$) for that replicate. This comparison was also performed for one replicate for each of the scenarios to evaluate the impact of increasing the number of individuals compared to the number of markers. Finally, for Scenarios 1 and 3, the widths of the 95% credible sets obtained from the posterior distributions of accuracy across all replicates was compared to the widths of the 95% confidence intervals of the accuracy of the genetic merit to evaluate whether the posterior distributions of accuracy were capturing variation in accuracy estimates due to sampling of training and validation individuals as well as uncertainty in marker effect estimates.

4.4 Results

Posterior Distribution of Accuracy

Figure 4.2 shows the posterior distribution of accuracy across all MCMC samples for a single replicate of Scenario 1. The mean of the posterior distributions of the accuracies ($\overline{r_{(y,u^*)}}$ (Figure 4.2A), $\overline{r_{(u,u^*)}}$ (Figure 4.2B) and $\overline{\tilde{r}_{(u,u^*)}}$ (Figure 4.2C)) were lower than the point estimates of accuracies of the posterior estimates of genetic merit ($r_{(y,\hat{u})}$ (Figure 4.2A), $r_{(u,\hat{u})}$ (Figure 4.2B) and $\tilde{r}_{(u,\hat{u})}$ (Figure 4.2C)). Tables S4.1–S4.3 show the results of Figure 4.2 for five replicates of the simulation. Results were similar among different replicates.

Figure 4.2B shows the correlation between the true genetic merit of the individual (u) and either u* or uHat. The approximation of this correlation ($\tilde{r}_{(u,u^*)}$; (Figure 4.2C)) was found to be reasonable because the prediction accuracy based on uHat was approximately the same in

Figures 4.2B and 4.2C. The 95% credible set from the posterior distribution of accuracy in Figure 4.2C, the approximation, was slightly wider than the credible set from Figure 4.2B (Tables S4.2 and S4.3).

Impact of Number of Individuals vs. Number of Markers

The distance between the posterior mean accuracy between y and u* ($\overline{r_{(y,u^*)}}$) and the accuracy based on posterior estimates of breeding values ($r_{(y,\hat{u})}$) decreased as the number of individuals increased relative to the number of markers (Figure 4.3). This was also observed for the correlation between u and u* (Figure S4.4). Prediction accuracy increased and the spread of the posterior distribution of prediction accuracy decreased as the number of individuals increased relative to the number of markers (Figure 4.3 and S4.4).

Spread of the Posterior Distribution of Accuracy

Tables 4.2 and 4.3 compare the mean and 95% confidence interval ($\hat{u}$) or credible set (u*) across the 200 replicates of Scenario 1 or 160 replicates of Scenario 3, respectively. The average correlation for the posterior estimates of genetic merit ($r(\hat{u},\cdot)$) was larger than the average of the posterior mean prediction accuracy ($\overline{r(u^*,\cdot)}$). The width of the 95% confidence interval of $r(\hat{u},\cdot)$ represents the variance in the accuracy estimate that is due to changing the training set, validation set and due to the uncertainty in the marker effect estimates obtained from each training analysis. The width of the credible sets from the posterior distribution of accuracy is similar across replicates and the approximated correlation between u* and u ($\tilde{r}(u^*,u)$) has the largest variance in the width of the 95% credible set across replicates (Tables 4.2 and 4.3). When the number of individuals is less than the number of markers (Scenario 1, Table 4.2) the width of

the credible set is similar to the width of the confidence interval when considering the correlation with y and u (i.e. $r(\cdot, y)$ and $r(\cdot, u)$); however, the width of the confidence interval for the approximated correlation (i.e. $\tilde{r}(\cdot, u)$) was much larger than the 95% credible set. When the number of individuals is more than the number of markers (Scenario 3, Table 4.3) the width of the correlation with u (i.e. $r(\cdot, u)$) is the same as the width of the credible set from the posterior distribution of accuracy; however, the width of the confidence interval is higher than the widths of the credible sets for both $r(\cdot, y)$ and $\tilde{r}(\cdot, u)$. These results suggest that the posterior distribution of accuracy is appropriately capturing the uncertainty in our estimate of prediction accuracy when considering the correlation between u and $\hat{u}$ but not when considering the correlation between y and $\hat{u}$ or the approximated correlation between u and $\hat{u}$.

4.5 Discussion

Evaluating the Posterior Distribution of Accuracy

We were able to develop a method to obtain a distribution of accuracies from the samples of α obtained each iteration of the MCMC; however, this distribution was not centered around the accuracy of the posterior estimates of genetic merit in this simulated data set (Figures 4.2 and 4.3). The posterior mean of a function is only guaranteed to be the mean of the posterior distribution of that function if the function is linear, however the accuracy function is not linear (i.e. $\text{mean}\left(\frac{\text{var(u)}}{\text{var(y)}}\right)$ is not guaranteed to equal $\frac{\text{mean(var(u))}}{\text{mean(var(y))}}$). As the number of individuals in the training set increases relative to the number of markers the variance of u* will approach the variance of $\hat{u}$ and the mean of the posterior distribution of accuracy will approach the accuracy of the posterior estimate of genetic merit. This can be observed in Figure 4.3 where the posterior

distribution of accuracy gets closer to the accuracy of the posterior genetic merit estimates when the number of markers is reduced (Figure 4.3A vs. 4.3B) or when the number of individuals in the training data set is increased (Figure 4.3B vs. 4.3C).

The approach taken in this study is promising because uncertainty in the estimated correlation between the estimated and true genetic merit of an individual ($\hat{u}$ and u) was appropriately captured through the credible set from the posterior distribution of accuracy (Tables 4.2 and 4.3). The true genetic merit of an individual is not observed in practice, therefore this correlation is only able to be obtained in a simulation study. When considering the correlation between the estimated genetic merit and the phenotype of that individual ($\hat{u}$ and y) or when the correlation between the estimated and true genetic merit was approximated, the posterior distribution of accuracy underestimated the confidence in these accuracy estimates because the credible set is narrower than the confidence interval (Tables 4.2 and 4.3).

Future Directions

The simulation used was based on Karaman et al. (2016) with parameters selected in an attempt to make the findings broadly applicable to human, plant and livestock populations. The SNP density was chosen because the 10 chromosomes in this simulation cover 1 Morgan, so if scaled to a 30 Morgan genome, similar humans or cattle, would represent a 600,000 SNP panel. Panels with approximately 600,000 SNPs are available in humans (Illumina and Affymetrix), cattle (Matukumalli et al., 2011) chickens (Kranis et al., 2013) and maize (Unterseer et al., 2014). The heritability used for the simulation was 0.5 to enable the accurate estimation of genetic merit without the need for a very large training population and considering the estimated heritability for a number of traits that may benefit from genomic prediction. Traits with a

heritability of approximately 0.5 include intelligence in humans (Davies et al., 2011), cobb mass in maize (Flint-Garcia et al., 2005) and weight in cattle (Bullock et al., 1993; Kaps et al., 1999).

This study focused on one heritability and a limited number of different number of animals and markers. It is important to follow up our study by evaluating the relationship between the posterior distribution of accuracy and the true confidence in accuracy estimates from the posterior marker effects by simulating traits with different heritability, different numbers of markers and with greater or fewer QTL. Another important aspect to investigate is how the relationship between individuals impacts our results. Prediction accuracy is influenced by the relationships between individuals in the training and validation sets (Goddard, 2009). The data set simulated in this study included many half-siblings and some full-siblings because each parent had six offspring on average. Most genomic prediction studies in humans contain distantly related individuals, often of diverse ethnicities and the predictive ability in another population is of interest (Yang et al., 2010). It is expected that further investigation will reveal information about the relationship between the uncertainty in the accuracy estimate and the posterior distribution of accuracy that can be used to modify the approach that we have presented.

Practical Implications

The method we describe for calculating the posterior distribution of accuracy requires less computing effort and is therefore faster than using either bootstrapping or cross-validation methods. Unlike cross-validation or bootstrapping of training individuals, the genomic prediction model only needs to be run once for our method. Rather than needing to sample multiple validation datasets, as for bootstrapping, our method only requires the multiplication of two matrices (genotype matrix of validation individuals and the marker effect samples across

iterations) and the calculation of the correlation. This method can therefore be easily incorporated into current MCMC genomic prediction workflows.

The posterior distribution of accuracy would ideally be centered at the accuracy of the posterior estimates of genetic merit and further research should explore whether there is an equation that will center this distribution and whether that distribution appropriately measures our confidence in this estimate. Our study has compared the posterior distribution of accuracy to the distribution of accuracy of the posterior estimates of genetic merit across many replicates of our simulation. The replicates of our simulation capture variation due to training individuals, validation individuals and uncertainty in the model estimates. In practice, the training set is fixed based on the individuals that have genotypes and phenotypes, therefore the uncertainty in the estimate of accuracy may not need to capture variation due to sampling of training individuals. Only replicating the simulation of validation sets of individuals rather than both training and validation sets will likely bring the 95% credible set from the posterior distribution of accuracy more in line with the 95% confidence interval from accuracy calculated between $\hat{u}$ and y or the approximated correlation between $\hat{u}$ and u across replicates.

Conclusions

The development of a method to evaluate the posterior distribution of prediction accuracy will provide additional information that can be used to make more robust decisions about implementation of genomic selection if that posterior distribution appropriately captures the variation in accuracy estimates that are likely to be observed in practice. For example, in the case where it is only beneficial to implement genomic selection if prediction accuracy is above a certain threshold, it is possible to calculate the proportion of the posterior distribution that is

above the desired threshold. With the range of Bayesian genomic prediction models available it is common to evaluate the performance of different models on a given data set to determine the most accurate model to use. Comparison of an appropriate posterior distribution of accuracy for each model provides more information than only comparing the accuracy of the posterior estimates of genetic merit because one model may have a slightly lower accuracy but a much larger variance, while another model may have a higher accuracy with less variation. This information will enable more informed decisions as to the most appropriate model to use on a given data set for genomic prediction.

4.6 Acknowledgements

The authors would like to thank Dr Wan-Ling Nicole Hsu for her assistance with Julia and XSim and Dr Alicia Carriquiry for her insightful discussions on Bayesian analyses.

4.7 Author Contributions

MH ran the analyses, interpreted the results, and wrote the manuscript. AH assisted with interpretation of results. RF supervised the study. All authors contributed to study design and critically contributed to the manuscript.

4.8 References

Allen, H. L., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, T. Ferreira, A. R. Wood, R. J. Weyant, A. V. Segre, E. K. Speliotes, E. Wheeler, N. Soranzo, J. H. Park, J. Yang, D. Gudbjartsson, N. L. Heard-Costa, J. C. Randall, L. Qi, A. V. Smith, R. Magi, T. Pastinen, L. Liang, I. M. Heid, J. Luan, G. Thorleifsson, T. W. Winkler, M. E. Goddard, K. S. Lo, C. Palmer, T. Workalemahu, Y. S. Aulchenko, A. Johansson, M. C. Zillikens, M. F. Feitosa, T. Esko, T. Johnson, S. Ketkar, P. Kraft, M. Mangino, I. Prokopenko, D. Absher, E. Albrecht, F. Ernst, N. L. Glazer, C. Hayward, J. J. Hottenga, K. B. Jacobs, J. W. Knowles, Z. Kutalik, K. L. Monda, O. Polasek, M. Preuss, N. W. Rayner, N. R. Robertson, V. Steinthorsdottir, J. P. Tyrer, B. F. Voight, F. Wiklund, J. F. Xu, J. H. Zhao, D. R. Nyholt, N. Pellikka, M. Perola, J. R. B. Perry, I. Surakka, M. L. Tammesoo, E. L. Altmaier, N. Amin, T. Aspelund, T. Bhangale, G. Boucher, D. I. Chasman, C. Chen, L. Coin, M. N. Cooper, A. L. Dixon, Q. Gibson, E. Grundberg, K. Hao, M. J. Junttila, L. M. Kaplan, J. Kettunen, I. R. Konig, T. Kwan, R. W. Lawrence, D. F. Levinson, M. Lorentzon, B. McKnight, A. P. Morris, M. Muller, J. S. Ngwa, S. Purcell, S. Rafelt, R. M. Salem, E. Salvi, S. Sanna, J. X. Shi, U. Sovio, J. R. Thompson, M. C. Turchin, L. Vandenput, D. J. Verlaan, V. Vitart, C. C. White, A. Ziegler, P. Almgren, A. J. Balmforth, H. Campbell, L. Citterio, A. De Grandi, A. Dominiczak, J. Duan, P. Elliott, R. Elosua, J. G. Eriksson, N. B. Freimer, E. J. C. Geus, N. Glorioso, S. Haiqing, A. L. Hartikainen, A. S. Havulinna, A. A. Hicks, J. N. Hui, W. Igl, T. Illig, A. Jula, E. Kajantie, T. O. Kilpelaeinen, M. Koiranen, I. Kolcic, S. Koskinen, P. Kovacs, J. Laitinen, J. J. Liu, M. L. Lokki, A. Marusic, A. Maschio, T. Meitinger, A. Mulas, G. Pare, A. N. Parker, J. F. Peden, A. Petersmann, I. Pichler, K. H. Pietilainen, A. Pouta, M. Riddertrale, J. I. Rotter, J. G. Sambrook, A. R. Sanders, C. O. Schmidt, J. Sinisalo, J. H. Smit, H. M. Stringham, G. B. Walters, E. Widen, S. H. Wild, G. Willemsen, L. Zagato, L. Zgaga, P. Zitting, H. Alavere, M. Farrall, W. L. McArdle, M. Nelis, M. J. Peters, S. Ripatti, J. B. J. Meurs, K. K. Aben, K. G. Ardlie, J. S. Beckmann, J. P. Beilby, R. N. Bergman, S. Bergmann, F. S. Collins, D. Cusi, M. den Heijer, G. Eiriksdottir, P. V. Gejman, A. S. Hall, A. Hamsten, H. V. Huikuri, C. Iribarren, M. Kahonen, J. Kaprio, S. Kathiresan, L. Kiemeney, T. Kocher, L. J. Launer, T. Lehtimaki, O. Melander, T. H. Mosley, A. W. Musk, M. S. Nieminen, C. J. O'Donnell, C. Ohlsson, B. Oostra, L. J. Palmer, O. Raitakari, P. M. Ridker, J. D. Rioux, A. Rissanen, C. Rivolta, H. Schunkert, A. R. Shuldiner, D. S. Siscovick, M. Stumvoll, A. Tonjes, J. Tuomilehto, G. J. van Ommen, J. Viikari, A. C. Heath, N. G. Martin, G. W. Montgomery, M. A. Province, M. Kayser, A. M. Arnold, L. D. Atwood, E. Boerwinkle, S. J. Chanock, P. Deloukas, C. Gieger, H. Gronberg, P. Hall, A. T. Hattersley, C. Hengstenberg, W. Hoffman, G. M. Lathrop, V. Salomaa, S. Schreiber, M. Uda, D. Waterworth, A. F. Wright, T. L. Assimes, I. Barroso, A. Hofman, K. L. Mohlke, D. I. Boomsma, M. J. Caulfield, L. A. Cupples, J. Erdmann, C. S. Fox, V. Gudnason, U. Gyllensten, T. B. Harris, R. B. Hayes, M. R. Jarvelin, V. Mooser, P. B. Munroe, W. H. Ouwehand, B. W. Penninx, P. P. Pramstaller, T. Quertermous, I. Rudan, N. J. Samani, T. D. Spector, H. Volzke, H. Watkins, J. F. Wilson, L. C. Groop, T. Haritunians, F. B. Hu, R. C. Kaplan, A. Metspalu, K. E. North, D. Schlessinger, N. J. Wareham, D. J. Hunter, J. R. O'Connell, D. P. Strachan, H. E. Schadt, U. Thorsteinsdottir, L. Peltonen, A. G. Uitterlinden, P. M.

Visscher, N. Chatterjee, R. J. F. Loos, M. Boehnke, M. I. McCarthy, E. Ingelsson, C. M. Lindgren, G. R. Abecasis, K. Stefansson, T. M. Frayling, J. N. Hirschhorn, and C. Procardis. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467(7317):832-838. doi: 10.1038/nature09410

Altshuler, D. M., R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. M. Zhu, Y. Q. Chang, Q. Feng, X. D. Fang, X. S. Guo, M. Jian, H. Jiang, X. Jin, T. M. Lan, G. Q. Li, J. X. Li, Y. R. Li, S. M. Liu, X. Liu, Y. Lu, X. D. Ma, M. F. Tang, B. Wang, G. B. Wang, H. L. Wu, R. H. Wu, X. Xu, Y. Yin, D. D. Zhang, W. W. Zhang, J. Zhao, M. R. Zhao, X. L. Zheng, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M. L. Yaspo, L. Fulton, R. Fulton, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. L. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. P. Zhan, A. Auton, C. L. Campbell, Y. Kong, A. Marcketta, F. L. Yu, L. Antunes, M. Bainbridge, A. Sabo, Z. Y. Huang, L. J. M. Coin, L. Fang, Q. B. Li, Z. Y. Li, H. X. Lin, B. H. Liu, R. B. Luo, H. J. Shao, Y. L. Xie, C. Ye, C. Yu, F. Zhang, H. C. Zheng, H. M. Zhu, C. Alkan, E. Dal, F. Kahveci, E. P. Garrison, D. Kural, W. P. Lee, W. F. Leong, M. Stromberg, A. N. Ward, J. T. Wu, M. Y. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, E. Banks, G. Bhatia, G. del Angel, G. Genovese, H. Li, S. Kashin, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, T. Rausch, M. H. Fritz, A. M. Stuetz, K. Beal, A. Datta, J. Herrero, G. R. S. Ritchie, D. Zerbino, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, B. Barnes, M. Bauer, R. K. Cheetham, A. Cox, M. Eberle, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Herwig, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. F. Willems, J. T. Simpson, M. D. Shriver, J. A. Rosenfeld, C. D. Bustamante, S. B. Montgomery, F. M. De La Vega, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. S. Zhang, F. C. L. Hyland, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, C. L. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. J. Kent, B. Howie, A. Ruiz-

Linares, E. T. Dermitzakis, S. E. Devine, R. A. Goncalo, H. M. Kang, J. M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. S. Li, R. Lyons, C. Scheller, C. Sidore, S. Y. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. K. Wing, X. W. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. H. Shi, A. Quitadamo, G. Lunter, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. X. Fu, X. M. Liu, M. M. Xiong, L. Jorde, D. Witherspoon, J. C. Xing, B. L. Browning, S. R. Browning, F. Hormozdiari, P. H. Sudmant, E. Khurana, C. Tyler-Smith, C. A. Albers, Q. Ayub, Y. Chen, V. Colonna, L. Jostins, K. Walter, Y. L. Xue, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. M. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, G. Del Angel, C. Hartl, K. Shakir, J. Degenhardt, S. Meiers, B. Raeder, F. P. Casale, O. Stegle, E. W. Lameijer, I. Hall, V. Bafna, J. Michaelson, E. J. Gardner, R. E. Mills, G. Dayama, K. Chen, X. Fan, Z. C. Chong, T. H. Chen, M. J. Chaisson, J. Huddleston, M. Malig, B. J. Nelson, N. F. Parrish, B. Ben, S. J. Lindsay, Z. M. Ning, Y. J. Zhang, H. Lam, C. Sisu, D. Challis, U. S. Evani, J. Lu, U. Nagaswamy, J. Yu, W. S. Li, L. Habegger, H. Y. Yu, F. Cunningham, I. Dunham, K. Lage, J. B. Jespersen, H. Horn, D. Kim, R. Desalle, A. Narechania, M. A. W. Sayres, F. L. Mendez, G. D. Poznik, P. A. Underhill, L. Coin, D. Mittelman, R. Banerjee, M. Cerezo, T. Fitzgerald, S. Louzada, A. Massaia, G. R. Ritchie, F. T. Yang, D. Kalra, W. Hale, X. Dan, K. C. Barnes, C. Beiswanger, H. Y. Cai, H. Z. Cao, B. Henn, D. Jones, J. S. Kaye, A. Kent, A. Kerasidou, R. Mathias, P. N. Ossorio, M. Parker, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Y. Su, Z. M. Tian, S. Tishkoff, M. Via, Y. H. Wang, H. M. Yang, L. Yang, J. Y. Zhu, W. Bodmer, G. Bedoya, Z. M. Cai, Y. Gao, J. Y. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, X. Y. Deng, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. S. Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. T. Hien, S. J. Dunstan, N. T. Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L. D. Brook, A. Felsenfeld, J. E. McEwen, Y. Vaydylevich, A. Duncanson, M. Dunn, J. A. Schloss, L. D. Brooks, and C. Genomes Project. 2015. A global reference for human genetic variation. Nature 526(7571):68-+. doi: 10.1038/nature15393

Bullock, K. D., J. K. Bertrand, and L. L. Benyshek. 1993. Genetic and Environmental Parameters for Mature Weight and Other Growth Measures in Polled Hereford Cattle. Journal of Animal Science 71(7):1737-1741.

Cheng, H., D. Garrick, and R. Fernando. 2015. XSim: Simulation of Descendants from Ancestors with Sequence Data. G3-Genes Genomes Genetics 5(7):1415-1417. doi: 10.1534/g3.115.016683

Davies, G., A. Tenesa, A. Payton, J. Yang, S. E. Harris, D. Liewald, X. Ke, S. Le Hellard, A. Christoforou, M. Luciano, K. McGhee, L. Lopez, A. J. Gow, J. Corley, P. Redmond, H. C. Fox, P. Haggarty, L. J. Whalley, G. McNeill, M. E. Goddard, T. Espeseth, A. J. Lundervold, I. Reinvang, A. Pickles, V. M. Steen, W. Ollier, D. J. Porteous, M. Horan, J. M. Starr, N. Pendleton, P. M. Visscher, and I. J. Deary. 2011. Genome-wide association

studies establish that human intelligence is highly heritable and polygenic. Molecular Psychiatry 16(10):996-1005. doi: 10.1038/mp.2011.85

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013a. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193(2):327-+. doi: 10.1534/genetics.112.143313

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen. 2013b. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. Plos Genetics 9(7)doi: 10.1371/journal.pgen.1003608

Flint-Garcia, S. A., A. C. Thuillet, J. M. Yu, G. Pressoir, S. M. Romero, S. E. Mitchell, J. Doebley, S. Kresovich, M. M. Goodman, and E. S. Buckler. 2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant Journal 44(6):1054-1064. doi: 10.1111/j.1365-313X.2005.02591.x

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257. doi: 10.1007/s10709-008-9308-0

Goddard, M. E. 2001. The validity of genetic models underlying quantitative traits. Livestock Production Science 72(1-2):117-127. doi: 10.1016/s0301-6226(01)00272-x

Gray, I. C., D. A. Campbell, and N. K. Spurr. 2000. Single nucleotide polymorphisms as tools in human genetics. Human Molecular Genetics 9(16):2403-2408. doi: 10.1093/hmg/9.16.2403

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389-2397. (Article) doi: 10.1534/genetics.107.081190

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. Bmc Bioinformatics 12:12. (Article) doi: 10.1186/1471-2105-12-186

Hess, M., T. Druet, A. Hess, and D. Garrick. 2016. Fixed length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Submitted to GSE*

Kaps, M., W. O. Herring, and W. R. Lamberson. 1999. Genetic and environmental parameters for mature weight in Angus cattle. Journal of Animal Science 77(3):569-574.

Karaman, E., H. Cheng, M. Firat, D. Garrick, and R. Fernando. 2016. An upper bound for accuracy of prediction using GBLUP. PLOS One

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. Journal of Animal Science 88(2):544-551. (Article) doi: 10.2527/jas.2009-2064

Kranis, A., A. A. Gheyas, C. Boschiero, F. Turner, L. Yu, S. Smith, R. Talbot, A. Pirani, F. Brew, P. Kaiser, P. M. Hocking, M. Fife, N. Salmon, J. Fulton, T. M. Strom, G. Haberer, S. Weigend, R. Preisinger, M. Gholami, S. Qanbari, H. Simianer, K. A. Watson, J. A. Woolliams, and D. W. Burt. 2013. Development of a high density 600K SNP genotyping array for chicken. Bmc Genomics 14doi: 10.1186/1471-2164-14-59

Matukumalli, L., S. Schroeder, S. DeNise, T. Sonstegard, C. Lawley, M. Georges, W. Coppieters, K. Gietzen, J. Medrano, and G. Rincon. 2011. Analyzing LD blocks and CNV segments in cattle: novel genomic features identified using the BovineHD BeadChip. San Diego, CA: Illumina Inc

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. Plos One 4(4)doi: 10.1371/journal.pone.0005350

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829. (Article)

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. M. Dekkers, R. L. Fernando, R. D. Schnabel, D. J. Garrick, and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics Selection Evolution 43doi: 10.1186/1297-9686-43-40

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2008. Genomic selection using different marker types and densities. Journal of Animal Science 86(10):2447-2454. doi: 10.2527/jas.2007-0010

Spichenok, O., Z. M. Budimlija, A. A. Mitchell, A. Jenny, L. Kovacevic, D. Marjanovic, T. Caragine, M. Prinz, and E. Wurmbach. 2011. Prediction of eye and skin color in diverse populations using seven SNPs. Forensic Science International-Genetics 5(5):472-478. doi: 10.1016/j.fsigen.2010.10.005

Unterseer, S., E. Bauer, G. Haberer, M. Seidel, C. Knaak, M. Ouzunova, T. Meitinger, T. M. Strom, R. Fries, H. Pausch, C. Bertani, A. Davassi, K. F. X. Mayer, and C. C. Schon. 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. Bmc Genomics 15doi: 10.1186/1471-2164-15-823

Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). (Accessed 10 November 2016.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M.

Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42(7):565-U131. (Article) doi: 10.1038/ng.608

Zeng, J. 2015. Whole genome analyses accounting for structures in genotype data, Iowa State University, http://lib.dr.iastate.edu/etd/14699.

4.8 Tables and Figures

Table 4.1: Posterior distribution of accuracy (u*) and the distribution of the accuracy of the genetic merit estimates ($\hat{u}$) across 200 replicates for Scenario 1[1]

Correlation[2]	$r(\hat{u},\cdot)$	Mean $r(u^*,\cdot)$			Width of 95% CI $r(\hat{u},\cdot)$[3]	Width of 95% Credible Set $r(u^*,\cdot)$		
		Minimum	Mean	Maximum		Minimum	Mean	Maximum
$r(\cdot,y)$	0.583	0.436	0.493	0.563	0.089	0.056	0.062	0.068
$r(\cdot,u)$	0.829	0.660	0.701	0.738	0.048	0.050	0.058	0.064
$\tilde{r}(\cdot,u)$	0.833	0.590	0.704	0.813	0.188	0.079	0.089	0.101

1) Scenario 1 simulates a training population of 5,000 individuals and validation set of 1,000 individuals for 20,000 markers
2) $r(\cdot,y)$ is the correlation between either u* or $\hat{u}$ and y; $r(\cdot,u)$ is the correlation between either u* or $\hat{u}$ and u; $\tilde{r}(\cdot,u)$ is the approximated correlation between either u* or $\hat{u}$ and u
3) CI = Confidence interval

Table 4.2: Posterior distribution of accuracy (u*) with the distribution of the accuracy of the genetic merit estimates ($\hat{u}$) across 160 replicates for Scenario 3[1]

Correlation[2]	$r(\hat{u},\cdot)$	Mean $r(u^*,\cdot)$			Width of 95% CI $r(\hat{u},\cdot)$[3]	Width of 95% Credible Set $r(u^*,\cdot)$		
		Minimum	Mean	Maximum		Minimum	Mean	Maximum
$r(\cdot,y)$	0.783	0.726	0.760	0.790	0.049	0.018	0.020	0.022
$r(\cdot,u)$	0.962	0.924	0.934	0.944	0.011	0.010	0.011	0.013
$\tilde{r}(\cdot,u)$	0.969	0.840	0.940	1.045	0.134	0.022	0.024	0.027

1) Scenario 3 simulates a training population of 20,000 individuals and validation set of 1,000 individuals for 10,000 markers
2) $r(\cdot,y)$ is the correlation between either u* or $\hat{u}$ and y; $r(\cdot,u)$ is the correlation between either u* or $\hat{u}$ and u; $\tilde{r}(\cdot,u)$ is the approximated correlation between either u* or $\hat{u}$ and u
3) CI = Confidence interval

Supplemental Table S4.1: Correlation between y and $\hat{u}$ and the Mean, Median and 95% Credible Set of Correlation between y and u*

Replicate	r(y, $\hat{u}$)	r(y,u*)		
		Mean	Median	95% Credible Set
1	0.58	0.48	0.48	(0.45,0.52)
2	0.53	0.45	0.45	(0.42,0.48)
3	0.61	0.52	0.52	(0.49,0.55)
4	0.60	0.51	0.51	(0.48,0.54)
5	0.58	0.49	0.49	(0.46,0.52)

Supplemental Table S4.2: Correlation between u and $\hat{u}$ and the Mean, Median and 95% Credible Set of Correlation between u and u*

Replicate	r(u, $\hat{u}$)	r(u,u*)		
		Mean	Median	95% Credible Set
1	0.81	0.68	0.68	(0.65,0.71)
2	0.81	0.68	0.68	(0.65,0.71)
3	0.83	0.71	0.71	(0.68,0.73)
4	0.82	0.70	0.70	(0.67,0.73)
5	0.84	0.71	0.71	(0.68,0.74)

Supplemental Table S4.3: Estimated Correlation between u and $\hat{u}$ and the Mean, Median and 95% Credible Set of Correlation between u and u*

Replicate	$\tilde{r}$(u, $\hat{u}$)	$\tilde{r}$ (u,u*)		
		Mean	Median	95% Credible Set
1	0.81	0.68	0.68	(0.64,0.73)
2	0.77	0.65	0.65	(0.60,0.69)
3	0.85	0.73	0.73	(0.69,0.77)
4	0.86	0.73	0.73	(0.68,0.77)
5	0.80	0.67	0.67	(0.63,0.71)

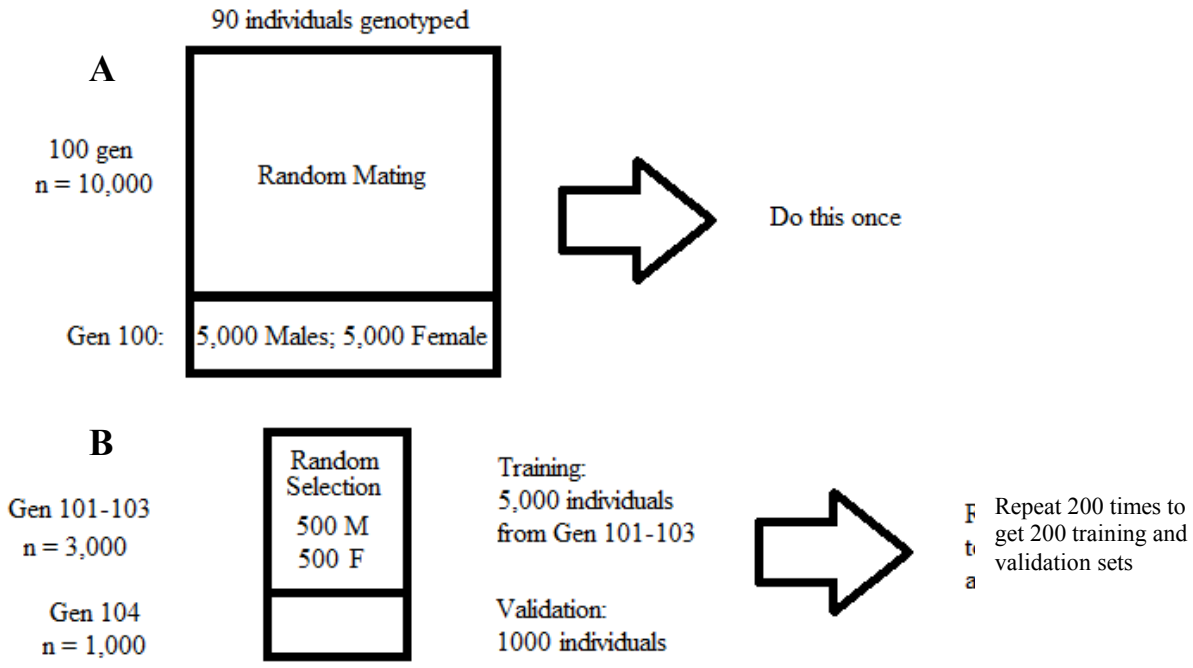


Figure 4.1: Simulation of Training and Validation Scenario 1 Data Sets

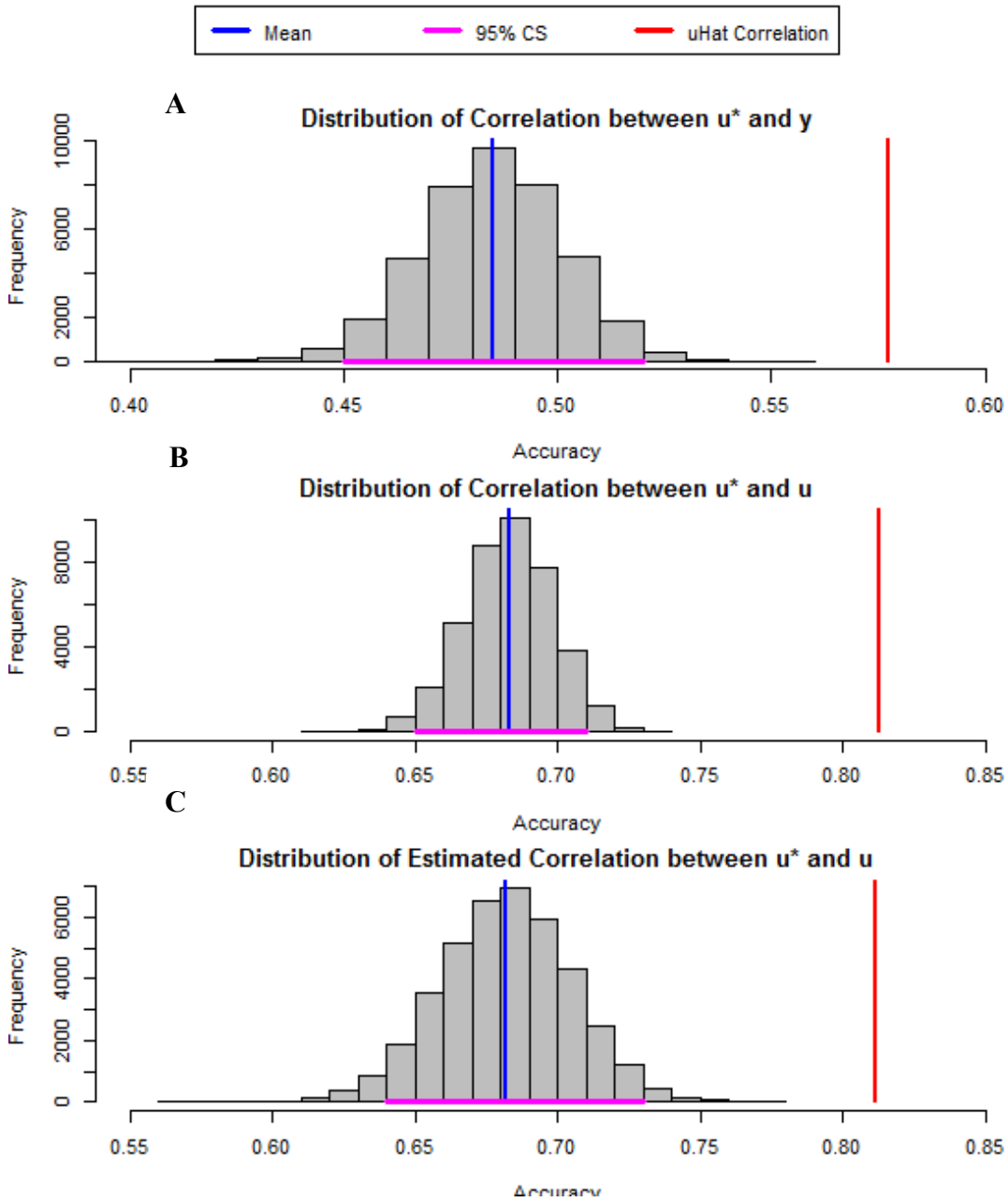


Figure 4.2: Posterior Distribution of Accuracy Compared to the Accuracy of Posterior Estimates. The genetic merit of an individual from each sample from the MCMC is u* and the posterior estimate of the genetic merit of an individual is uHat ($\hat{u}$). The phenotypic value is y and the equation for the estimated correlation can be found in equation 1. The blue line is the mean of the posterior distribution, the red line is the accuracy of the posterior breeding values and the pink line represents the 95% credible set (CS) of the posterior distribution.

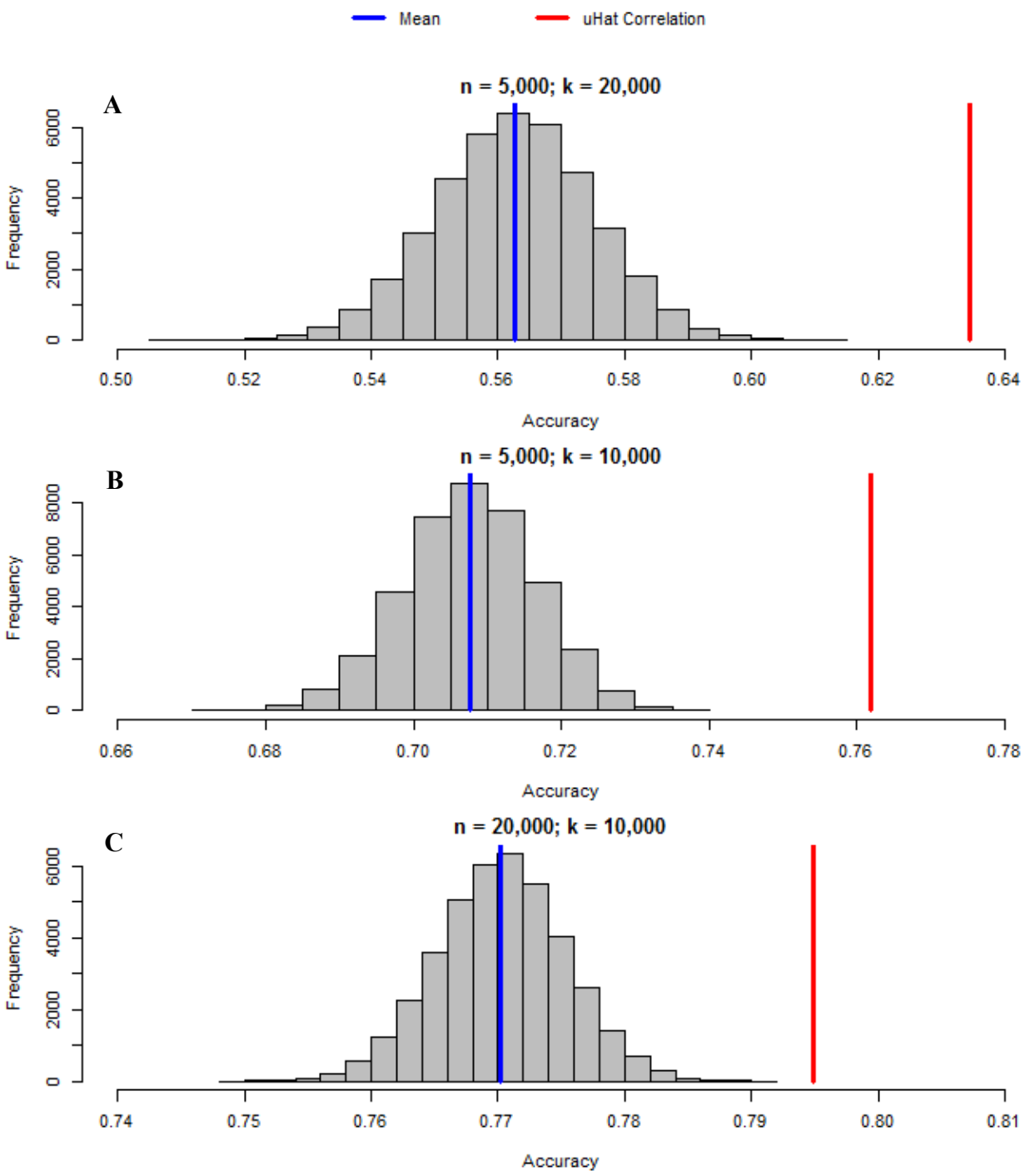
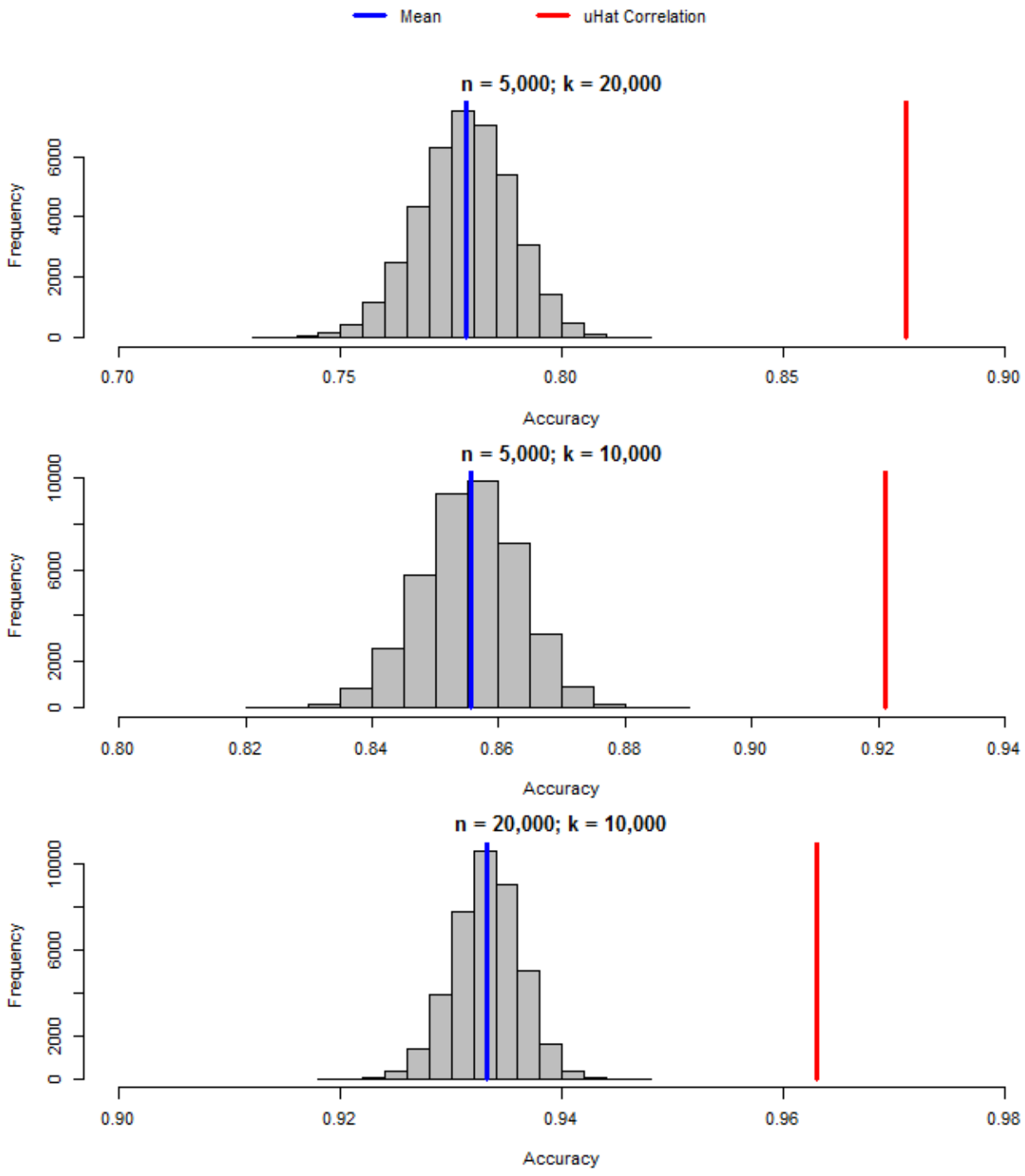


Figure 4.3: Posterior Distribution of Correlation between y and u* (Blue) and Correlation between y and $\hat{u}$ (Red) with Varying Numbers of Training Individuals (n) and Markers (k).



Supplemental Figure S4.4: Posterior Distribution of Correlation between u and u* (Blue) and Correlation between u and $\hat{u}$ (Red) with Varying Numbers of Training Individuals (n) and Markers (k).

CHAPTER V

GENERAL DISCUSSION

Genomic prediction accuracy and bias have a direct impact on genetic gain in populations that undergo genomic selection and many resources are dedicated to the development of models that improve prediction accuracy (Lush, 1937; Meuwissen et al., 2013). Genomic prediction models fitting haplotype alleles rather than SNPs may improve prediction accuracy if haplotypes have stronger linkage disequilibrium (LD) with QTL than SNPs, or if haplotypes are better than SNPs at capturing the realized genetic relationship between individuals, i.e. the proportion of the genome that is identical-by-decent (IBD).

5.1 Research Objectives

The overall objective of this dissertation was to improve accuracy of genomic prediction in the admixed New Zealand dairy cattle population by fitting covariates for haplotype alleles rather than covariates for SNPs. Improved accuracy of genomic prediction should translate to increased rates of genetic gain. Chapter II explored the use of fixed-length haplotype alleles from 125 kb to 2 Mb in length, the application of four levels of haplotype allele frequency filters, and the performance of three Bayesian genomic prediction models with different prior assumptions for allele effect estimates. Chapter III compared the performance of genomic prediction fitting variable-length haplotype alleles based on pairwise or multi-locus linkage disequilibrium (LD), recombination events, or reducing the

number of haplotype alleles. Chapter IV investigated an approach to obtain a posterior distribution of prediction accuracy that appropriately captures the variation in prediction accuracy that may be observed in practice and could be used for more reliable comparison of the accuracy of genomic prediction models.

5.2 Prediction Accuracy

Performance of genomic prediction models is typically evaluated based on their prediction accuracy and bias (Daetwyler et al., 2013). These are evaluated by separating the set of genotyped and phenotyped animals into training and validation sets, with the intent that estimates from the validation set represent the accuracy and bias in the un-phenotyped selection candidates. When comparing two models to evaluate their potential for improving rates of genetic gain, it is important to obtain an understanding of the confidence in the accuracy estimates because one may have a slightly higher accuracy in the validation set but a larger standard error, and therefore not give consistently high accuracy when applied to the selection candidates.

Two common methods for obtaining a standard error of the accuracy estimate are cross-validation (Saatchi et al., 2011) and bootstrapping (Cuyabano et al., 2015a). One of the possible advantages of haplotype models is that they better capture co-ancestry between individuals than SNP models (Ferdosi et al., 2016). Cross-validation is not a recommended approach for comparing prediction accuracy when comparing the use of haplotypes versus SNPs because the relationship between individuals in the training set and validation set may not appropriately model the relationships between the full set of genotyped and phenotyped individuals and future selection candidates and, thus, may underestimate the advantage of

using haplotypes. The posterior distribution of prediction accuracy (Chapter IV) keeps the training and validation sets separate, so the older individuals can be used for training and the younger animals can be used for validation, mimicking the relationships between the full set of individuals and future selection candidates. The approach that was taken in Chapter IV attempted to capture variation in accuracy estimates due to different groups of training and validation individuals, as well as from uncertainty in marker effect estimates. In practice, it may be more appropriate to capture variation only due to uncertainty in marker effects estimates and in validation animals because the training data set is usually fixed by the individuals in the population that have genotypes and phenotypes. The bootstrap approach applied to the validation set of animals allows for paired t-tests to be performed by sampling a validation set of individuals and comparing accuracy (or bias) in each of the two models that are being compared – and repeating this process several (e.g. 10000) times; this allows for a more powerful comparison than a non-paired t-test across the same number of samples. It is not possible to pair samples of the posterior accuracy across two models because each sample is independent; therefore it is not appropriate to pair them.

Future Directions

The approach in Chapter IV, where we developed a method to obtain the posterior distribution of accuracy deserves further investigation. It is an efficient method for evaluating the confidence in an estimate of prediction accuracy because it doesn't necessitate running the prediction model multiple times, as in cross-validation or bootstrapping of training individuals; and, unlike the bootstrapping of validation individuals, uncertainty in marker effects estimates is captured. The next steps for research into this approach should include

evaluation of the distribution in different simulated scenarios such as traits with different heritabilities and populations with different mating scenarios. It is also important to gain an understanding of the relationship between the mean of the posterior distribution of accuracy and the accuracy of posterior estimates of marker effects.

5.3 Haplotype Analyses

Haploblock Methods

Improvements in prediction accuracy were observed when fitting haplotype alleles rather than SNPs using either fixed-length (Chapter II; up to 5.5% improvement: $p < 0.003$) or variable-length haploblocks (Chapter III; up to 7.7% improvement: $p < 0.001$). Bias was not significant when fitting SNPs for Holstein Friesian or KiwiCross cows but the regression of yield deviation on genomic breeding value was greater than one for Jerseys, which made the top individuals appear worse than they were. In some cases, fitting haplotype alleles decreased bias compared to the SNP model in Jerseys but there were no significant decreases in bias in Holstein Friesian or KiwiCross cows. Therefore, in addition to improving accuracy, haplotype models may also reduce bias.

Fixed-Length Haploblocks

The length of fixed-length haploblocks has been shown to impact prediction accuracy in many studies (Hayes et al., 2007; Calus et al., 2008; Calus et al., 2009; Villumsen and Janss, 2009; Villumsen et al., 2009; Ferdosi et al., 2016). Chapter II showed that smaller haploblocks of 125 – 250 kb tended to improve genomic prediction accuracy compared to the SNP model in this population, however when haploblocks were longer than 0.5 Mb, prediction accuracy decreased and bias increased. The removal of haplotype alleles with

frequencies less than 10% in the training population had a much larger impact on longer haploblocks than shorter haploblocks because of the increase in the number of rare haplotype alleles as haploblocks get longer. The appropriate length of haploblock depends on marker density, as well as on population parameters such as LD; therefore, it is important to identify the optimal haploblock length independently for each data set (Calus et al., 2009; Villumsen et al., 2009). Fixed-length haploblocks have benefits over variable-length haploblocks in that they are more straightforward to generate and remain unchanged over many generations, assuming the same SNP panel is used. Haploblocks of 125 kb and removing haplotype alleles with frequencies less than 10% in the training data set fitted a similar number of haplotype alleles as the SNP model, and therefore had comparable runtime, but improved or maintained accuracy compared to the SNP model (Chapter II). Therefore, fitting haplotype alleles generated from fixed-length haploblocks is a straightforward method that generally improves or maintains genomic prediction accuracy in purebred and admixed populations, given that haploblocks lengths used are appropriate for the population.

Variable-Length Haploblocks

Variable-length haploblocks attempt to capture the true haploblock structure of the population better than fixed-length haploblocks, and the haploblocks generated from these methods are likely specific to that population. Chapter III evaluated the performance of genomic prediction using four variable-length haploblock methods.

Methods based on LD did not consistently show improvement over SNP or fixed-length haploblocks as had been expected. This may be due to the admixed-breed nature of the data set used in this study; patterns of LD are specific to breed (de Roos et al., 2009) and, although marker phase has been shown to be reasonably conserved between New Zealand

Holstein Friesians and Jerseys (de Roos et al., 2008) measurements of LD calculated across the whole population may not appropriately capture LD patterns within each breed. Therefore, LD-based methods may perform better in more homogenous populations, where they have previously been shown to improve prediction accuracy (Cuyabano et al., 2015a).

Haploblocks that were generated using identified recombination events within the population had the greatest improvement in prediction accuracy of any of the haplotype models evaluated as part of this dissertation; closely followed by the method that aimed to reduce the number of haplotype alleles fitted in the model (Chapter III). The performance of the recombination model may be sensitive to the ability to accurately identify recombination events in the population and therefore may not always perform better than the method that aims to reduce the number of haplotype alleles. The recombination method to generate haploblocks will perform best when power to detect recombination events is high (i.e. when a large number of animals are genotyped, in particular many related individuals). The data used in this dissertation is an example where the recombination method is likely to perform well because there were >36,000 parent-offspring pairs that were genotyped, allowing the accurate identification of recombination events. If recombination events cannot be accurately identified, the haploblock method that aims to reduce the number of haplotype alleles is an attractive approach because its accuracy was usually similar to the accuracy of the recombination method. One downfall of this method is that it is more computationally-intensive than the recombination method to generate the haploblocks (most of the computational power needed to construct recombination-based haploblocks occurs during phasing); however once the haploblocks have been generated there are fewer haplotype

alleles to fit, so the genomic prediction model runs faster than when haplotype alleles are generated from recombination haploblocks.

Reducing Dimensionality of Haplotype Analyses

An issue with haplotype analyses is that the number of haplotype alleles can be much larger than the number of SNPs (Hayes et al., 2007). When using high-density (~777,000) SNP genotypes and creating haploblocks based on LD, Cuyabano et al. (2015a) identified ~318,000 haplotype alleles, which was fewer than the number of SNPs, so they did not attempt to reduce the number of covariates further. Other studies performing genomic prediction using haplotypes either ignored this dimensionality issue (Pryce et al., 2010) or addressed it by fitting haplotypes in only some genomic regions (Boichard et al., 2012) or by removing low-minor-allele-frequency SNPs prior to haplotype construction (Calus et al., 2009; Cuyabano et al., 2015b). None of these studies reported on the impact of their approach to reducing the number of covariates on prediction accuracy or bias.

The studies in this dissertation reduced the number of covariates in genomic prediction models fitting haplotype alleles by removing rare haplotype alleles (Chapter II and III), which is similar to the common approach of removing SNPs with low minor allele frequency when performing genomic prediction using SNPs (VanRaden et al., 2009; Harris and Johnson, 2010; Hayes et al., 2010). An allele frequency threshold of 10% was found to reduce prediction accuracy compared to a threshold of 1% in fixed-length haplotypes, particularly when haploblocks were greater than 0.5 Mb in length (Chapter II). Applying a haplotype allele frequency threshold of 1% had similar prediction accuracy as fitting all haplotype alleles for variable-length haploblocks but only fitted 50 to 60% of the covariates

and therefore had faster computation time (Chapter III). The exception was the Multi-Locus LD haploblocks method, which had higher prediction accuracy when the 1% filter was applied, compared to fitting all haplotype alleles (Chapter III), which can be attributed to heavy shrinkage of effects for those rare alleles (Gianola, 2013), such that their estimates are very close to zero. Therefore, removing haplotype alleles with frequency less than 1% in the training population is a recommended approach to reducing the runtime of genomic prediction when fitting genome-wide haplotype alleles without sacrificing prediction accuracy. Prediction accuracy may be improved if the effects of rare alleles can be estimated based on similarity to more common haplotype alleles.

Bayesian Genomic Prediction Models

Chapter II showed that prediction accuracy was similar between methods BayesA, BayesB and BayesN when fitting covariates for fixed-length haplotype alleles of either 250 or 125 kb. These results were consistent across traits, although Milk Fat Yield and Liveweight both have large QTL (Grisart et al., 2002; Cohen-Zinder et al., 2005; Karim et al., 2011; Komisarek et al., 2011; Littlejohn et al., 2014) and Somatic Cell Score is a highly polygenic trait (Meredith et al., 2012). Chapter III only evaluated the performance of BayesA since Chapter II showed similar performance of the Bayesian models evaluated. However, it is possible that an alternative model, such as BayesN, would result in higher prediction accuracy when haploblocks model the recombination events and LD in the population rather than fixed-length haploblocks if the haploblocks capture more biologically-relevant information.

5.4 Future Directions

Genotyping in the Future

More Animals

As part of the now-routine genomic selection process in many dairy cattle populations, the majority of sires are genotyped, and focus has shifted to increasing the number of genotyped dams (Stock and Reents, 2013). Increasing the number of genotyped animals in genomic prediction models that fit SNP covariates will improve confidence in the estimated effects of each SNP and generally improve prediction accuracy (de los Campos et al., 2013; Hayes et al., 2009; Knol et al., 2016). Although the confidence in SNP estimates will be higher, increasing the number of genotyped animals used in genomic prediction will not directly account for context-specific QTL, such as those with breed-specific or genotype-by-environment effects; however increasing the number of animals will give more power to be able to detect these associations when they are appropriately modelled.

Increasing the number of genotyped animals may improve the accuracy of genomic prediction models that fit haplotype alleles in two ways. First, similar to genomic prediction fitting SNPs, increasing the number of individuals will improve confidence in estimates of the haplotype allele effects (assuming the number of individuals with that haplotype allele increases). A further improvement that is likely to come as a consequence of genotyping more individuals is the increased phasing accuracy (Browning and Browning, 2011), which will likely improve the ability to identify biologically-relevant haploblocks and correctly assign haplotype alleles to each individual. For example, when considering recombination-based haplotypes, a phasing error may identify a recombination event when there is none – potentially impacting haploblock boundaries; if the phasing error occurs within a designated

haploblock, that individual will be assigned two incorrect haplotype alleles for that haploblock, which may reduce prediction accuracy for that individual. Therefore, prediction accuracy is likely to improve as the number of individuals increase because we will have more power to define haploblocks and call haplotype alleles. Recently, a method to sequence haplotypes rather than genotypes was developed, which would allow for direct haplotyping of an individual (Noyes et al., 2015). This may result in higher accuracy of haploblock construction, which may be used in place of, or in concert with, phasing algorithms.

Markers

In addition to increased numbers of individuals being genotyped, there is also an increase in the number of approaches for genotyping those individuals. Commercially available SNP panels for dairy cattle range in density from ~3,000 up to ~777,000 SNPs (Stock and Reents, 2013). The cost of whole-genome sequence continues to fall and more than 3,000 dairy cattle have now been sequenced, with millions of genetic variants identified (Taylor, 2016). Recently, SNP panels have been produced that contain putative functional variants (Taylor, 2016; Taylor et al., 2016) or lower-density panels that have high imputation accuracy (Wu et al., 2016). Genomic prediction by fitting sequence variants was predicted to greatly improve prediction accuracy because the QTL would be included in the set of SNPs, rather than relying on LD to capture QTL effects (Meuwissen et al., 2013). In practice, these expected improvements in prediction accuracy have not been observed, likely because these models contain many more covariates than individuals or because sequencing is typically performed on a small subset of animals and the remainder are imputed up to sequence from SNP genotypes, resulting in imputation inaccuracies (van Binsbergen et al., 2015; Heidaritabar et al., 2016). Haplotypes generated from sequence variants or (putative)

functional variants may have the ability to detect interactions between these functional variants, i.e. short-range epistasis, when these variants are included in the same haploblock. The number of unique haplotype alleles will likely be less than the number of variants, so haplotype analyses are an appealing approach to incorporate sequence information into genomic prediction models. Low-density SNP panels with high imputation accuracy (Wu et al., 2016) could be used to track the inheritance of chromosome segments in females, which could then be imputed to a higher density to generate haplotype alleles for genomic prediction.

Modeling Haplotypes

Chapter III evaluated the performance of variable-length haploblocks based on a specific number of haploblocks, i.e. the number of haploblocks that performed the best for fixed-length haploblocks in Chapter II. It is probable that altering the number of haploblocks produced using variable-length methods would result in different estimates of prediction accuracy given the variation in recombination rates and LD across the cattle genome (Sandor et al., 2012; Weng et al., 2014). Chapter III also proposed adjustments to the haploblock methods that should be explored in the future, such as weighting the haplotype alleles by their frequency in the population when generating haploblocks that reduce the number of haplotype alleles.

Boichard et al. (2012) described the application of a combined SNP and haplotype genomic prediction model to the French dairy cattle evaluation program, where haplotype alleles are fitted around known or putative QTL (from a SNP model) and SNPs are fitted in the remainder of the genome. Although Boichard et al. (2012) observed improved accuracy

from fitting this model, other options for fitting a combined SNP and haplotype genomic prediction model should be explored. One approach could be to identify genomic regions that explain more of the genetic variance for a trait when fitting haplotype alleles rather than SNPs in genomic prediction models. This approach is likely to improve prediction accuracy compared to the approach by Boichard et al. (2012) because haplotype models likely capture the effects of QTL that are not captured in SNP models (Sun, 2014). In principle, this could be done by running a SNP model and a haplotype model then identifying the regions that explain more genetic variance under the haplotype model and only fit haplotype alleles in these regions and either SNPs in the remainder of the genome or a polygenic effect. Alternatively, a single Bayesian mixture model with both SNPs and haplotype alleles could be used.

Single-step genomic prediction models combine the analysis of genotyped and non-genotyped individuals into a single evaluation through a matrix that combines both marker- and pedigree-information (Legarra et al., 2009; Fernando et al., 2014). As discussed in Chapter I, the SNP-based **G** matrix and the **A** matrix are typically on different scales because they capture different information (identity-by-state (IBS) and IBD, respectively); therefore, current single-step GBLUP methods differentially weight the **A** and **G** matrices (Harris et al., 2012). However, the weighting used impacts prediction accuracy (Gao et al., 2012). If the **G** matrix is generated using haplotype information rather than SNP information, it is more similar to the **A** matrix than when it is generated using SNPs (Ferdosi et al., 2016; Hickey et al. 2013), provided appropriate length haplotypes are used. Thus, if haplotype information is used to generate the **G** matrix, differential weighting of the **A** and **G** matrices may not be required, or prediction accuracy estimates may be more robust to different weightings.

Therefore, it is expected that single-step methods using haplotypes will show an improvement over SNP-based single-step methods using SNPs because they better capture the realized genetic relationship between individuals.

5.5 Conclusions

Genomic prediction models that fit covariates for haplotype alleles rather than SNPs show promise for improvement of genetic gain in admixed populations, such as New Zealand dairy cattle. Genomic prediction accuracy depended on the method used to generate haplotypes and the optimal haplotypes to use for genomic prediction will likely be population-specific. Fixed-length haplotypes of 125 to 250 kb either maintained or improved prediction accuracy compared to genomic prediction based on SNPs. The haploblock method that utilized information on recombination events within the population had the highest prediction accuracy of the methods evaluated, closely followed by the method that reduced the number of haplotype alleles to fit in the genomic prediction model. Performance of the recombination method is expected to be highly dependent on the ability to accurately identify recombination events and therefore may not perform well in small data sets or those with few family members genotyped. Haploblock methods based on measurements of LD may not be appropriate for use in admixed populations because the LD patterns likely differ between the breeds that make up the population. The number of covariates to fit in genomic prediction models was successfully reduced by removing rare haplotype alleles with no cost to prediction accuracy. There are multiple avenues for research into haplotypes that could lead to further improvements in prediction accuracy including using functional variants to

improve the ability to identify short-range epistatic interactions; or developing models that fit haplotypes in some genomic regions and SNPs in the rest of the genome.

5.6 References

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Animal Production Science 52(2-3):115-120. (Review) doi: 10.1071/an11119

Browning, S. R., and B. L. Browning. 2011. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics 12(10):703-714. doi: 10.1038/nrg3054

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178(1):553-561. (Article) doi: 10.1534/genetics.107.080838

Calus, M. P. L., T. H. E. Meuwissen, J. J. Windig, E. F. Knol, C. Schrooten, A. L. J. Vereijken, and R. F. Veerkamp. 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. Genetics, Selection, Evolution 41(11):(15 January 2009). (article)

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, A. Everts-van der Wind, J. H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, H. A. Lewin, J. I. Weller, and M. Ron. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15(7):936-944. (Article) doi: 10.1101/gr.3806705

Cuyabano, B. C. D., G. Su, G. J. M. Rosa, M. S. Lund, and D. Gianola. 2015a. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. Journal of Dairy Science 98(10):7351-7363. doi: 10.3168/jds.2015-9360

Cuyabano, B. C. D., G. S. Su, and M. S. Lund. 2015b. Selection of haplotype variables from a high-density marker map for genomic prediction. Genetics Selection Evolution 47:11. (Article) doi: 10.1186/s12711-015-0143-3

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193(2):347-+. doi: 10.1534/genetics.112.147983

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193(2):327-+. doi: 10.1534/genetics.112.143313

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183(4):1545-1553. doi: 10.1534/genetics.109.104935

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179(3):1503-1512. (Article) doi: 10.1534/genetics.107.084301

Ferdosi, M. H., J. Henshall, and B. Tier. 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genetics Selection Evolution

Fernando, R. L., J. C. M. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution 46doi: 10.1186/1297-9686-46-50

Gao, H. D., O. F. Christensen, P. Madsen, U. S. Nielsen, Y. Zhang, M. S. Lund, and G. S. Su. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genetics Selection Evolution 44doi: 10.1186/1297-9686-44-8

Gianola, D. 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194(3):573-596. (Article) doi: 10.1534/genetics.113.151753

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12(2):222-231. (Article) doi: 10.1101/gr.224202

Harris, B. L., A. M. Winkelman, and D. L. Johnson. 2012. Large-scale single-step genomic evaluation for milk production traits No. 46. Interbull Bulletin.

Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. Journal of Dairy Science 93(3):1243-1252. doi: 10.3168/jds.2009-2619

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92(2):433-443. (Review) doi: 10.3168/jds.2008-1646

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetics Research 89(4):215-220. (Article) doi: 10.1017/s0016672307008865

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. Plos Genetics 6(9)doi: 10.1371/journal.pgen.1001139

Hickey, J. M., B. P. Kinghorn, B. Tier, S. A. Clark, J. H. J. van der Werf, and G. Gorjanc. 2013. Genomic evaluations using similarity between haplotypes. Journal of Animal Breeding and Genetics 130(4):259-269. doi: 10.1111/jbg.12020

Heidaritabar, M., M. P. L. Calus, H. J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics

Karim, L., H. Takeda, L. Lin, T. Druet, J. A. C. Arias, D. Baurain, N. Cambisano, S. R. Davis, F. Farnir, B. Grisart, B. L. Harris, M. D. Keehan, M. D. Littlejohn, R. J. Spelman, M. Georges, and W. Coppieters. 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nature Genetics 43(5):405-+. (Article) doi: 10.1038/ng.814

Knol, E. F., B. Nielsen, and P. W. Knap. 2016. Genomic selection in commercial pig breeding. Animal Frontiers 6(1):15-22.

Komisarek, J., A. Michalak, and A. Walendowska. 2011. The effects of polymorphisms in DGAT1, GH and GHR genes on reproduction and production traits in Jersey cows. Animal Science Papers and Reports 29(1):29-36. (Article)

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92(9):4656-4663. doi: 10.3168/jds.2009-2061

Littlejohn, M. D., K. Tiplady, T. Lopdell, T. A. Law, A. Scott, C. Harland, R. Sherlock, K. Henty, V. Obolonkin, K. Lehnert, A. MacGibbon, R. J. Spelman, S. R. Davis, and R. G. Snell. 2014. Expression Variants of the Lipogenic AGPAT6 Gene Affect Diverse Milk Composition Phenotypes in Bos taurus. Plos One 9(1):12. (Article) doi: 10.1371/journal.pone.0085757

Lush, J. L. 1937. Animal breeding plans. Animal breeding plans.:Pp. x + 350. (Book)

Meredith, B. K., F. J. Kearney, E. K. Finlay, D. G. Bradley, A. G. Fahey, D. P. Berry, and D. J. Lynn. 2012. Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. Bmc Genetics 13doi: 10.1186/1471-2156-13-21

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating Improvement of Livestock with Genomic Selection. In: H. A. Lewin and R. M. Roberts, editors, Annual Review of Animal Biosciences, Vol 1. Annual Review of Animal Biosciences No. 1. Annual Reviews, Palo Alto. p. 221-237.

Noyes, H. A., D. Daly, I. Goodhead, S. Kay, S. J. Kemp, J. Kenny, I. Saccheri, R. D. Schnabel, J. F. Taylor, and N. Hall. 2015. A simple procedure for directly obtaining haplotype sequences of diploid genomes. Bmc Genomics 16doi: 10.1186/s12864-015-1818-4

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J. Bowman, K. Savin, M. E. Goddard, and B. J. Hayes. 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. Journal of Dairy Science 93(7):3331-3345. doi: 10.3168/jds.2009-2893

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. M. Dekkers, R. L. Fernando, R. D. Schnabel, D. J. Garrick, and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics Selection Evolution 43doi: 10.1186/1297-9686-43-40

Sandor, C., W. B. Li, W. Coppieters, T. Druet, C. Charlier, and M. Georges. 2012. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. Plos Genetics 8(7):13. (Article) doi: 10.1371/journal.pgen.1002854

Stock, K. F., and R. Reents. 2013. Genomic Selection: Status in Different Species and Challenges for Breeding. Reproduction in Domestic Animals 48:2-10. doi: 10.1111/rda.12201

Sun, X. 2014. Genomic prediction using linkage disequilibrium and co-segregation, Iowa State University, http://lib.dr.iastate.edu/etd/14273.

Taylor, J. F. 2016. Design and Application of the Cattle GGP-F250 Array. In *Plant and Animal Genome XXIV Conference*. Plant and Animal Genome. In: Plant and Animal Genome XXIV Conference, San Diego, CA

Taylor, J. F., L. K. Whitacre, J. L. Hoff, P. C. Tizioto, J. Kim, J. E. Decker, and R. D. Schnabel. 2016. Lessons for livestock genomics from genome and transcriptome sequencing in cattle and other mammals. Genetics Selection Evolution 48doi: 10.1186/s12711-016-0237-6

van Binsbergen, R., M. P. L. Calus, M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47:13. (Article) doi: 10.1186/s12711-015-0149-x

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92(1):16-24. doi: http://dx.doi.org/10.3168/jds.2008-1514

Villumsen, T. M., and L. Janss. 2009. Bayesian genomic selection: the effect of haplotype length and priors. BMC proceedings 3 Suppl 1:S11.

Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics 126(1):3-13. (Article) doi: 10.1111/j.1439-0388.2008.00747.x

Weng, Z. Q., M. Saatchi, R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genetics Selection Evolution 46doi: 10.1186/1297-9686-46-34

Wu, X. L., J. Q. Xu, G. F. Feng, G. R. Wiggans, J. F. Taylor, J. He, C. S. Qian, J. S. Qiu, B. Simpson, J. Walker, and S. Bauck. 2016. Optimal Design of Low-Density SNP Arrays for Genomic Prediction: Algorithm and Applications. Plos One 11(9)doi: 10.1371/journal.pone.0161719

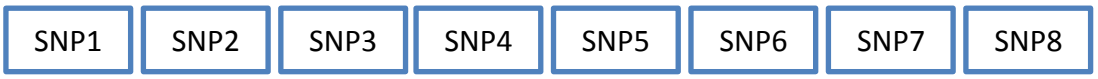
APPENDIX A

ALGORITHM FOR GENERATING HAPLOBLOCKS

This appendix will go through a small example of the algorithm that was used to assign SNPs to haploblocks using the Pairwise LD method described in Chapter III.

We have **eight SNPs**: SNP1 – SNP8; and want to assign them into **five haploblocks**.

Step 1: Assign Each SNP to a unique haploblock



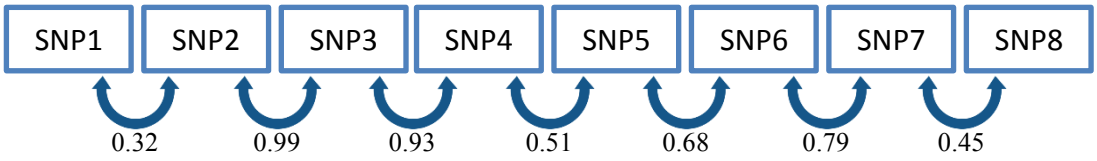
Step 2: Calculate Measurement Between Haploblocks

Measurement: Minimum D' between SNPs in neighboring blocks

At this stage there is only one SNP per block so measurement is just D'

$$D' = \frac{Pr(A_1B_1) - Pr(A_1)Pr(B_1)}{D_{max}}$$

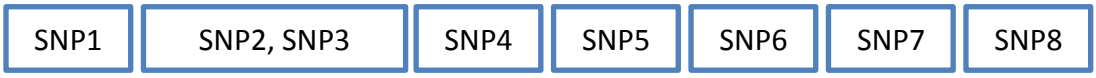
$$D_{max} = \begin{cases} Min\big(Pr(A_1)Pr(B_1), Pr(A_2)Pr(B_2)\big), if\ D < 0 \\ Min\big(Pr(A_1)Pr(B_2), Pr(A_2)Pr(B_1)\big), if\ D > 0 \end{cases}$$



Step 3: Determine Which Haploblocks to Join

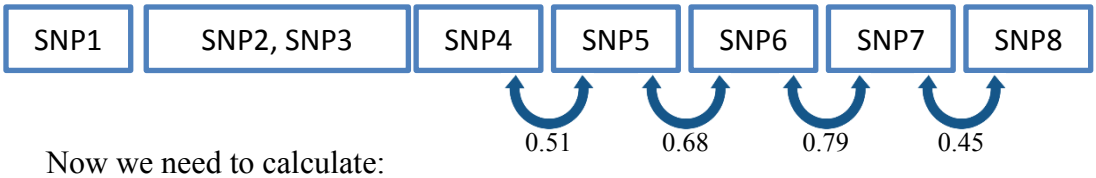
Joining Criteria: Maximum

So join **SNP2 and SNP3** (Measurement = 0.99)



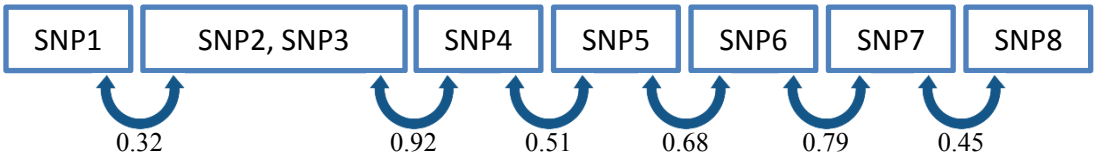
Step 4: Calculate Measurement Between New and Neighboring Haploblocks

We already have the measurement calculated for these blocks:

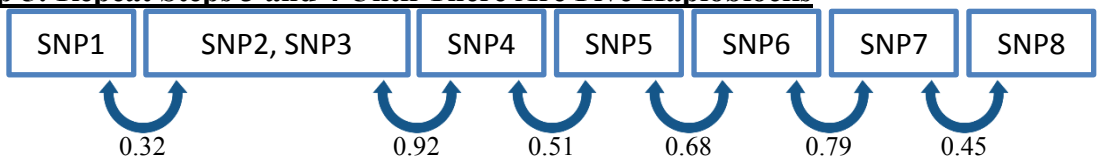


Now we need to calculate:

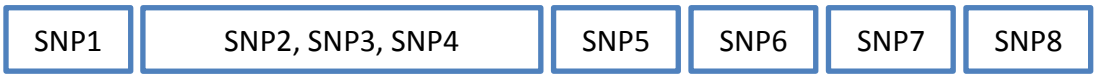
$\text{Min}(D'_{SNP1,SNP2}, D'_{SNP1,SNP3})$ and $\text{Min}(D'_{SNP2,SNP4}, D'_{SNP3,SNP4})$



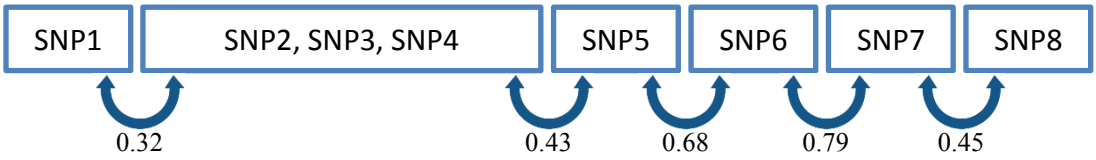
Step 5: Repeat Steps 3 and 4 Until There Are Five Haploblocks



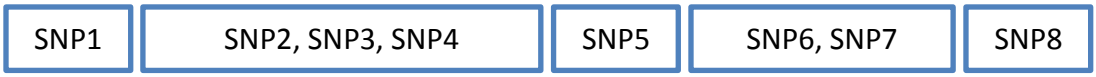
Join the haploblock of SNP2, SNP3 with the SNP4 haploblock:



Get the Measurement for new haploblocks:



Join the SNP6 haploblock with SNP7 haploblock:



The eight SNPs have been assigned to five haploblocks!

SNP	Haploblock
1	1
2	2
3	2
4	2
5	3
6	4
7	4
8	5

APPENDIX B

IDENTIFICATION OF RECOMBINATION EVENTS

Step 1: Determining Which Offspring Strand was Inherited from that Parent

It is easy to tell which offspring strand is from that parent by comparing the ancestral haplotypes (an output of LINKPHASE3) of each offspring strand to the genotyped parent:

Offspring 1 = first strand from the offspring Offspring 2 = second strand from the offspring
Parental 1 = first strand from the parent Parental 2 = second strand from the parent

Ancestral Haplotypes:

SNP Number	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Offspring 1:	1	1	3	3	3	7	7	2	2	2
Offspring 2:	7	7	7	7	5	5	5	5	5	5
Parental 1:	1	1	3	3	3	3	3	3	3	3
Parental 2:	9	9	9	7	7	7	7	2	2	2

Comparison of Strands (i.e. O1 == P1 for comparing O1 and P1):

SNP Number	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
O1 and P1:	1	1	1	1	1	0	0	0	0	0
O1 and P2:	0	0	0	0	0	1	1	1	1	1

Sum = 10

SNP Number	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
O2 and P1:	0	0	0	0	0	0	0	0	0	0
O2 and P2:	0	0	0	1	0	0	0	0	0	0

Sum = 1

Conclusion: The first strand (Offspring 1) is the strand inherited from the genotyped parent.

Step 2: Identifying Recombination Events

In some cases it is very clear where the recombination events occur, while in other places it is ambiguous. The clear presence/absence of recombination is shown in yellow and the ambiguous section is shown in grey. Under each example is the conclusion for each situation.
Case 1: A clear crossover between two SNPs

SNP Number	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
O1 and P1:	1	1	1	1	1	0	0	0	0	0
O1 and P2:	0	0	0	0	0	1	1	1	1	1

Conclusion: A recombination event occurred between SNP5 and SNP6

Case 2: Parent is homozygous for a stretch of SNPs and Parental strand 1 is the same as offspring strand before and after this stretch

SNP Number	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
O1 and P1:	1	1	1	1	1	1	1	1	1	1
O1 and P2:	0	0	0	1	1	1	1	1	0	0

Conclusion: No Recombination – it is possible there was a double recombination but it was assumed there was none.

Case 3: Parent is homozygous for a stretch of SNPs but the parental strand is different before and after this stretch

SNP Number	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
O1 and P1:	1	1	1	1	1	1	0	0	0	0
O1 and P2:	0	0	0	1	1	1	1	1	1	1

Conclusion: Recombination could be between SNPs 3-4, 4-5, 5-6, or 6-7. Each possibility was given a weighting of 0.25 because there were 4 options.

Step 3: Number of Recombination Events between Pairs of SNPs

The recombination events identified in Step 2 were then summed across all parent-offspring pairs to get the number of recombination events between each pair of consecutive SNPs across the genome.

APPENDIX C

SIMULATION OF THE BASE POPULATION

Packages needed

In [1]:
```
using XSim
using Distributions
using StatsBase
```

Function to Write Out Haplotype Data

In [2]:
```
function getHapData(this::XSim.Cohort)
    numAnimals = length(this.animalCohort)
    numLoci = sum([XSim.common.G.chr[chr].numLoci for chr in 1:XSim.common.G
    haps = convert(Array{Any,2},zeros(numAnimals*2,numLoci+1))
    pos = "AID"
    an = 1
    for animal in this.animalCohort
        haps[an*2-1,1] = string(animal.myID) * "_pat"
        haps[an*2,1] = string(animal.myID) * "_mat"
        loc = 1
        for chr in 1:XSim.common.G.numChrom
            for locus in 1:XSim.common.G.chr[chr].numLoci
                if (an == 1)
                    pos = [pos string(chr)*"_"*string(round(XSim.common.G.ch
                end
                haps[an*2-1,loc+1] = animal.genomePat[chr].haplotype[locus]
                haps[an*2,loc+1] = animal.genomeMat[chr].haplotype[locus]
                loc = loc+1
            end
        end
        an = an+1
    end
    return pos, haps
end;
```

Part One: Random Mating from Human 1000 Genomes Data

Map and Genotype Information

Read in marker information and separate by chromosome

In [3]:
```
genotypeFile = "markers.header"
genos = readdlm(genotypeFile,header=false)

chr = [0 for i in 2:length(genos)]
posTemp = [0.0 for i in 2:length(genos)];

for s in 2:length(genos)
    ss = split(genos[s],"_")
    chr[s-1] = parse(Int64,ss[1])
    posTemp[s-1] = parse(Float64,ss[2])
end
```

Get Map and Position Information into Correct Format

In [4]:
```
chr = chr .- minimum(chr) .+ 1
p = posTemp[find(chr .== 1)]
p = (p .- minimum(p))/100
nL = length(p)
pos = p
nChr = length(unique(chr))
if nChr > 1
    ch = unique(chr[chr.!=1])
    for c in 1:length(ch)
        p = posTemp[find(chr .== ch[c])]
        p = (p .- minimum(p))/100
        pos = [pos;p]
        nL = [nL;length(p)]
    end
    chrEnd = cumsum(nL)
    chrStart = [1;cumsum(nL)[1:end-1]+1]
    pos2 = tuple([pos[chrStart[c]:chrEnd[c]] for c in 1:nChr]...)
    gF = tuple([fill(0.5,nL[c]) for c in 1:nChr]...)
    qM = tuple([[true;fill(false,nL[c])] for c in 1:nChr]...)
    qE = tuple([[randn();fill(0,nL[c])] for c in 1:nChr]...);
else
    chrEnd = nL
    chrStart = 1
    pos2 = pos
    gF = fill(0.5,nL)
    qM = [true;fill(false,nL)]
    qE = [randn();fill(0,nL)]
end;
```

Parameters for the Simulation

Random mating of a population of 10,000 individuals for 100 generations

In [5]:
```
randomMatingSize = 10000
randomMatingGens = 100

numChr = nChr
chrLength = ifelse(numChr == 1,0.1,fill(0.1,numChr))
numLoci = nL
mutRate = 0.0
mapPos = pos2
geneFreq = gF
qtlMarker  = qM
qtlEffects = qE
XSim.init(numChr,numLoci,chrLength,geneFreq,mapPos,qtlMarker,qtlEffects,mutR
```

Randomly Select Founders (Generation 1)

In [6]:
```
popSizeFounder = convert(Int64,round(randomMatingSize/2))
founders = sampleFounders(90,"markers.haps")
sires = XSim.cohortSubset(founders,collect(1:45));
dams  = XSim.cohortSubset(founders,collect(46:90));
```

```
Sampling 90 animals into base population.
```

Random Mating Generations

In [7]:
```
sires1,dams1,gen1 = sampleRan(randomMatingSize, randomMatingGens-1, sires, c
```

```
Generation    95: sampling  5000 males and  5000 females
Generation    96: sampling  5000 males and  5000 females
Generation    97: sampling  5000 males and  5000 females
Generation    98: sampling  5000 males and  5000 females
Generation    99: sampling  5000 males and  5000 females
Generation   100: sampling  5000 males and  5000 females
```

Sample SNPs and QTL

Select Markers to Use as SNPs and QTL

- Combine male and female cohorts from Generation 100
- Filter Markers based on MAF > 0.005
- Randomly select markers from the good markers to be SNPs and QTL

In [8]:
```
mafCutoff = 0.005
baseCohort = concatCohorts(sires1,dams1)
XSim.getOurHaps(baseCohort)
pos,haps = getHapData(baseCohort);
```

In [9]:
```
a = [baseCohort.animalCohort[i].myID for i in 1:length(baseCohort.animalCohc
genos = [a getOurGenotypes(baseCohort)]
afs = sum(haps[:,2:end],1)/size(haps)[2]
mafs = [minimum([afs[i],1-afs[i]]) for i in 1:length(afs)]
keep = [1;find(mafs .> mafCutoff)+1]
goodMarkers = haps[:,keep]
goodGenos = genos[:,keep]
goodPos = pos[1,keep]
println("Number of Good Markers: "*string(length(keep)-1))
```

```
Number of Good Markers: 49509
```

Specify number of SNPs and QTL

In [10]:
```
numSNPs = 20000
numQTL = 300;
```

Randomly select markers and print results files

In [11]:
```
if((numSNPs+numQTL)< (size(goodMarkers)[2]-1))
    selMark = [1;sort(sample(1:(size(goodMarkers)[2]-1),numSNPs+numQTL,repla
    selHaps =  goodMarkers[:,selMark]
    selGens =  goodGenos[:,selMark]
    selPos = goodPos[1,selMark]
    qtlMarker  = fill(false,1,length(selPos)-1)
    qtlMarker[1,sample(1:length(qtlMarker),numQTL,replace=false)]= true
    qtlEffects = fill(0.0,1,length(qtlMarker))
    qtlEffects[1,qtlMarker] = randn(numQTL)
    writedlm("RandomMating.QTL",qtlMarker," ")
    writedlm("RandomMating.QTLEff",qtlEffects," ")
    writedlm("RandomMating.haps",selHaps," ")
    writedlm("RandomMating.genos",selGens," ")
    writedlm("RandomMating.header",selPos," ")
    gS = convert(Array{Float64,2},selGens[:,2:end])
    geneticVariance = var(gS*qtlEffects')
    println("Genetic Variance = "*string(round(geneticVariance,2)))
    writedlm("RandomMating.genVar",round(geneticVariance,4))
else
    println("Not enough markers")
end
```

```
Genetic Variance = 110.58
```

APPENDIX D

SIMULATION OF TRAINING AND VALIDATION DATA SETS

Packages Needed

In [1]:
```
using XSim
using StatsBase
```

Part Two: Generating Training and Validation Data Sets

Read in the genotype and marker information from Part One

In [2]:
```
genotypeFile = "RandomMating.header"
genos = readdlm(genotypeFile,header=false)'
chr = [0 for i in 2:length(genos)]
posTemp = [0.0 for i in 2:length(genos)];
for s in 2:length(genos)
    ss = split(genos[s],"_")
    chr[s-1] = parse(Int64,ss[1])
    posTemp[s-1] = parse(ss[2])
end

qtl = readdlm("RandomMating.QTL",header=false)
qtlE = readdlm("RandomMating.QTLEff",header=false)

chr = chr .- minimum(chr) .+ 1
p = posTemp[find(chr .== 1)]
p = p .- minimum(p)
nL = length(p)
pos = p
nChr = length(unique(chr))
if nChr > 1
    ch = unique(chr[chr.!=1])
    for c in 1:length(ch)
        p = posTemp[find(chr .== ch[c])]
        p = p .- minimum(p)
        pos = [pos;p]
        nL = [nL;length(p)]
    end
    chrEnd = cumsum(nL)
    chrStart = [1;cumsum(nL)[1:end-1]+1]
    pos2 = tuple([pos[chrStart[c]:chrEnd[c]] for c in 1:nChr]...)
    gF = tuple([fill(0.5,nL[c]) for c in 1:nChr]...)
    qM = tuple([qtl[chrStart[c]:chrEnd[c]] for c in 1:nChr]...)
    qE = tuple([qtlE[chrStart[c]:chrEnd[c]] for c in 1:nChr]...);
else
    chrEnd = nL
    chrStart = 1
    pos2 = pos
    gF = fill(0.5,nL)
    qM = convert(Array{Bool,1},reshape(qtl,nL))
    qE = reshape(readdlm("RandomMating.QTLEff",header=false),nL)
end;
```

Simulation Parameters

Heritability of 0.5

Training population of three generations of 3000 individuals with 500 males and 500 females as parents of the next generation. 5000 of these individuals are selected to be part of the genotyped and phenotyped training set.

Validation population is the next generation after the training population. 1000 individuals are selected to be genotyped and phenotyped and make up the validation set.

In [3]:
```
h2 = 0.5

trainPopSize = 3000
trainPopGens = 3
trainNumSires = 500
trainNumDams = 500
trainSampleSize = 5000

valPopSize = 3000
valPopGens = 1
valNumSires = 500
valNumDams = 500
valSampleSize = 1000;
```

In [4]:
```
geneticVariance = readdlm("RandomMating.genVar",header=false)[1,1]
residualVariance = ((1-h2)*geneticVariance)/h2
```

Out[4]: 110.5759

Initialize the simulation model

In [5]:
```
numChr = nChr
chrLength = ifelse(numChr == 1,0.1,fill(0.1,numChr))
numLoci = nL
mutRate = 0.0
mapPos = pos2
geneFreq = gF
qtlMarker  = qM
qtlEffects = qE
XSim.init(numChr,numLoci,chrLength,geneFreq,mapPos,qtlMarker,qtlEffects,mutR
```

Simulate Training Population

In [6]:
```
numParents = trainNumSires+trainNumDams
parents = sampleFounders(numParents,"RandomMating.haps")
sires = XSim.cohortSubset(parents,collect(1:trainNumSires));
dams  = XSim.cohortSubset(parents,collect((trainNumSires+1):numParents));
```

```
Sampling 1000 animals into base population.
```

In [7]:
```
sires1,dams1,gens=sampleRan(trainPopSize, 1, sires, dams,gen=0);
trainPop = concatCohorts(sires1,dams1);
```

```
Generation     1: sampling  1500 males and  1500 females
```

In [8]:
```
if trainPopGens > 1
    for i in 2:trainPopGens
        selSires = cohortSubset(sires1,sample(1:length(sires1.animalCohort),
        selDams = cohortSubset(dams1,sample(1:length(dams1.animalCohort),tra
        gensTemp = gens
        sires1,dams1,gens=sampleRan(trainPopSize, 1, selSires, selDams,gen=g
        trainPop = concatCohorts(trainPop,sires1,dams1)
    end
end
```

```
Generation     2: sampling  1500 males and  1500 females
Generation     3: sampling  1500 males and  1500 females
```

Simulate Validation Population

In [9]:
```
selSires = cohortSubset(sires1,sample(1:length(sires1.animalCohort),valNumSi
selDams = cohortSubset(dams1,sample(1:length(dams1.animalCohort),valNumDams,
gensTemp = gens
sires1,dams1,gens=sampleRan(valPopSize, 1, selSires, selDams,gen=gensTemp);
valPop = concatCohorts(sires1,dams1);
```

```
Generation     4: sampling  1500 males and  1500 females
```

Sample Training and Validation Animals

In [10]:
```
trainSample = cohortSubset(trainPop,sample(1:length(trainPop.animalCohort),t
valSample = cohortSubset(valPop,sample(1:length(valPop.animalCohort),valSamp
```

Genotype and Phenotype Information for Sampled Animals

In [11]:
```
XSim.getOurHaps(trainSample)
XSim.getOurHaps(valSample)

trainGenotypes = getOurGenotypes(trainSample)
valGenotypes = getOurGenotypes(valSample)

keep = [!qtl[i] for i in 1:length(qtl)]
trainSNPs = trainGenotypes[:,find(keep)]
valSNPs = valGenotypes[:,find(keep)]

trainBVs = trainGenotypes*qtlE'
valBVs = valGenotypes*qtlE'

trainPhenotypes = getOurPhenVals(trainSample,residualVariance)
valPhenotypes = getOurPhenVals(valSample,residualVariance);
```

Write simulated data out

In [ ]:
```
writedlm("train.X",trainSNPs)
writedlm("val.X",valSNPs)
writedlm("train.u",trainBVs)
writedlm("val.u",valBVs)
writedlm("train.y",trainPhenotypes)
writedlm("val.y",valPhenotypes)
```

APPENDIX E

SINGLE VALUE DECOMPOSITION GENOMIC PREDICTION MODEL

In [1]:
```
using Distributions
```

Read in the data set and center the genotype matrix

This is shown with a smaller example than simulated in Supplemental Material 7

In [2]:
```
X = readdlm("train.X",header=false);
XV = readdlm("val.X",header=false);
y = readdlm("train.y",header=false);
yV = readdlm("val.y",header=false);
u = readdlm("train.u",header=false);
uV = readdlm("val.u",header=false);
eV = yV - uV;

nTrain,nMark = size(X)
nVal = length(yV)

colMeans = ones(nTrain)*mean(X,1)
X = X-colMeans
println("NumSNPs = "*string(nMark)*", NumTrainAns = "string(nTrain)*", NumVa
```

```
NumSNPs = 5000, NumTrainAns = 3000, NumValAns = 1000
```

SVD Decomposition

$$M = UDV'$$

Where

- M is the m x n genotype matrix
- U is a matrix of size m x m
- D is the m x m diagonal matrix of eigenvalues (d)
- V is a matrix of size n x m

The matrix R can also be defined such that $R = UD$ and

$$R'R = D'U'UD$$

U is orthonormal (i.e. $U'U = I$) so

$$R'R = D'D = D^2$$

D is a diagonal matrix so D^2 is also a diagonal matrix where the diagonal can be calculated as d^2 which is less computationally intensive than multiplying 2 matrices

```
In [3]: @time U,d,V = svd(X);
        D = diagm(d);
        R = U*D;
```

```
75.358161 seconds (102.87 k allocations: 622.877 MB, 0.23% gc time)
```

Running the Model

The normal model we run for BayesC0 is:

$$y = \mu + M\alpha + e$$

With the SVD we can replace M with UDV':

$$y = \mu + UDV'\alpha + e$$

Substituting in $R = UD$ and $\beta = V'\alpha$:

$$y = \mu + R\beta + e$$

Full Conditional for $\mu|ELSE$:

Proportional to a Normal distribution with mean $\hat{\mu}$ and variance $\frac{\sigma_e^2}{n}$

Full Conditional for $\beta_j|ELSE$:

Proportional to a Normal distribution with mean $\frac{R'w}{d_j^2+\frac{\sigma_e^2}{\sigma_\beta^2}}$ and variance $\frac{\sigma_e^2}{d_j^2+\frac{\sigma_e^2}{\sigma_\beta^2}}$ where

$w = y - \mu - \sum_{l \neq j} R_l \beta_l$

Full Conditional for $\sigma_\beta^2|ELSE$:

Proportional to a scaled inverted chi-square with $\tilde{\nu}_\beta = \nu_\beta + k$, where k is the number of columns of U, and scale parameter $\frac{\sum_k \beta_j^2 + \nu_\beta S_\beta^2}{\tilde{\nu}_\beta}$

Full Conditional for $\sigma_e^2|ELSE$:

Proportional to a scaled inverted chi-square with $\tilde{\nu}_e = \nu_e + n$, where n is the number of training individuals, and scale parameter $\frac{e'e + \nu_e S_e^2}{\tilde{\nu}_e}$, where $e = y - \mu - R\beta$

Initializing variables

In [4]:
```
chainLength     = 41000
burnIn          =  1000
dfEffectVar     =     4
nuRes           =     4
varGenotypic    =  110
varResidual     =  110

muTrack = fill(0.0,chainLength)
vare = copy(varResidual)
vareTrack = fill(0.0,chainLength)
betaStar = fill(0.0,size(R)[2])
alphaTrack = fill(0.0,chainLength,nMark)
varaTrack = fill(0.0,chainLength);
varBeta = varGenotypic/(nMark*0.5)
varBetaTrack = fill(0.0,chainLength,nMark)
scaleVar    = varBeta*(dfEffectVar-2)/dfEffectVar
scaleRes    = varResidual*(nuRes-2)/nuRes;
```

In [5]:
```
mu = mean(y)
ycorr = y - mu
for iter in 1:chainLength
        ycorr = ycorr .+ mu
        mu = mean(ycorr) + randn(1)[1]*sqrt(vare/nTrain)
        ycorr = ycorr .- mu
        muTrack[iter] = mu

        ycorr = ycorr + R*betaStar
        lhs = (d.^2) + vare./(varBeta)
        sampleMean = (R'ycorr)./lhs
        betaStar = randn(length(betaStar)).*sqrt(vare./lhs) + sampleMean
        alphaTrack[iter,:] = V*betaStar
        ycorr = ycorr - R*betaStar

        vare = (((ycorr'ycorr)[1]+nuRes*scaleRes)/(rand(Chisq(nTrain+nuRes))
        vareTrack[iter] = vare

        varBeta = ((sum((betaStar.^2))+dfEffectVar*scaleVar)/(rand(Chisq(ler
        varBetaTrack[iter] = varBeta

        vara = var(R*betaStar)
        varaTrack[iter] = vara
end
```

Calculate $\hat{u}$ and u^* from the Samples

u* can easily be calculated by multiplying the genotype matrix for the valiation set by the samples of alpha in each of the iterations

In [6]:
```
alphaHats = mean(alphaTrack[(burnIn+1):end,:],1)'
uHats = XV*alphaHats
uStars = XV*convert(Array{Float64,2},alphaTrack[(burnIn+1):end,:]');
```

Print Correlations

In [7]:
```
rUStarY = convert(Array{Float64,1},[cor(yV,uStars[:,i])[1] for i in 1:size(u
println("Correlation between u* and y in validation")
println("   Mean = "*string(mean(corUStarY)))
println("   Median = "*string(median(corUStarY)))
println("   95% Credible Set = ("*string(quantile(corUStarY,0.025))*","*string
println("Correlation between uHat and y in validation = "*string(cor(uHats,yV)
println()

rUStarU = convert(Array{Float64,1},[cor(uV,uStars[:,i])[1] for i in 1:size(u
println("Correlation between u* and u in validation")
println("   Mean = "*string(mean(corUStarU)))
println("   Median = "*string(median(corUStarU)))
println("   95% Credible Set = ("*string(quantile(corUStarU,0.025))*","*string
println("Correlation between uHat and u in validation = "*string(cor(uHats,uV)
println()

rApUStarU = convert(Array{Float64,1},[cov(yV,uStars[:,i])[1]/sqrt(mean(varaT
println("Approximated Correlation between u* and u in validation")
println("   Mean = "*string(mean(corApUStarU)))
println("   Median = "*string(median(corApUStarU)))
println("   95% Credible Set = ("*string(quantile(corApUStarU,0.025))*","*stri
println("Approximated Correlation between uHat and u in validation = "*string(
println()
```

```
Correlation between u* and y in validation
   Mean = 0.5426833333299189
   Median = 0.5428042176262402
   95% Credible Set = (0.5153944988119628,0.5688005585465127)
Correlation between uHat and y in validation = [0.600165577391488]

Correlation between u* and u in validation
   Mean = 0.80069094409787
   Median = 0.8009477498841389
   95% Credible Set = (0.7752460874044514,0.8242620385414834)
Correlation between uHat and u in validation = [0.8855132090537768]

Approximated Correlation between u* and u in validation
   Mean = 0.7879821617798365
   Median = 0.7881576871061
   95% Credible Set = (0.7483584742713817,0.8259046597117032)
Approximated Correlation between uHat and u in validation = 0.87144701164
2942

Posterior Mean Genetic Variance = 10.357403822465509
```